Table of contents:

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S14–S36
DOI 10.1002/acr.20621
© 2011, American College of Rheumatology

Measures of Rheumatoid Arthritis Disease Activity

Patient (PtGA) and Provider (PrGA) Global Assessment of Disease Activity, Disease Activity Score (DAS) and Disease Activity Score With 28-Joint Counts (DAS28), Simplified Disease Activity Index (SDAI), Clinical Disease Activity Index (CDAI), Patient Activity Score (PAS) and Patient Activity Score-II (PASII), Routine Assessment of Patient Index Data (RAPID), Rheumatoid Arthritis Disease Activity Index (RADAI) and Rheumatoid Arthritis Disease Activity Index-5 (RADAI-5), Chronic Arthritis Systemic Index (CASI), Patient-Based Disease Activity Score With ESR (PDAS1) and Patient-Based Disease Activity Score Without ESR (PDAS2), and Mean Overall Index for Rheumatoid Arthritis (MOI-RA)

JACLYN K. ANDERSON,[1] LANI ZIMMERMAN,[2] LIRON CAPLAN,[3] AND KALEB MICHAUD[4]

INTRODUCTION

Numerous rheumatoid arthritis (RA) disease activity measurement tools are currently available for use. The psychometric data related to these tools have been published over the course of decades and across numerous journals. Consequently, the majority of this information remains inaccessible to practicing and academic rheumatologists alike. To facilitate the availability of this information, the American College of Rheumatology (ACR) RA Clinical Disease Activity Measures Working Group performed a systematic review and compiled the resulting data into an evidence report, which constitutes the majority of this publication. The following psychometric analysis is based on currently published literature with abstract publications excluded. More recent measurement tools with inadequately published data, such as the Global Arthritis Score, may ultimately prove to be equal or superior to those discussed in this article. Time to complete each measure was divided into 1) time for patient completion without assistance from providers; 2) time for provider completion, including the time required to score the measure; and 3) time required for laboratory studies. Our assessment of the Disease Activity Score, Disease Activity Score with 28-joint counts, and Rheumatoid Arthritis Disease Activity Index updates that of Fransen et al, who performed this exercise in 2003 (1).

All RA disease activity measurement tools discussed herein produce a single continuous index of disease activity and were chosen based on expert panel recommendations for use in the clinical setting as well as their current use in the research and clinical settings. Continuous composite indices producing a single score have an advantage over the interpretation of individual components of disease activity as they provide clinically meaningful and reliable estimates of disease activity with interpretation of multiple data points simultaneously and are more responsive to change than single items (2). Continuous indices are less susceptible to selection bias related to the reporting of a single measurement (3) and are also preferable for statistical analysis in studies, with the most commonly used composite indices able to quantify disease activity, even in the lower ends of the scales (4). Additionally, composite indices are recommended by many insurers and regulators to justify escalation of RA therapy (5,6). On the other hand, simple visual analog scales are widely used and may be the most feasible method for disease activity monitoring in clinical settings. Consequently, the patient global assessment of disease activity and provider global assessment of disease activity will be included in this report, even though these measures may factor in irreversible damage.

Supported in part by the American College of Rheumatology.

[1]Jaclyn K. Anderson, DO, MS: University of Nebraska Medical Center, Omaha, and Abbott Laboratories, Abbott Park, Illinois; [2]Lani Zimmerman, PhD, RN: University of Nebraska Medical Center, Omaha; [3]Liron Caplan, MD, PhD: Denver Veterans Affairs Medical Center, Denver, Colorado, and University of Colorado School of Medicine, Aurora; [4]Kaleb Michaud, PhD: University of Nebraska Medical Center, Omaha, and National Data Bank for Rheumatic Diseases, Wichita, Kansas.

Dr. Anderson owns stock and/or holds stock options in Abbott Laboratories.

Address correspondence to Jaclyn K. Anderson, DO, MS, 200 Abbott Park Road, Department R4NE, Building AP34-1, Abbott Park, IL 60064-6187. E-mail: Jaclyn.Anderson@abbott.com.

Submitted for publication January 25, 2011; accepted in revised form May 9, 2011.

A number of measures are not included in this article for various reasons, including lack of use and clinic feasibility; the full list of starting measures and the methods for exclusion are covered in our parallel study. Despite inclusion of a continuous score, the ACR hybrid also requires scoring a change from a prior assessment and will not be discussed, as it is unlikely to be useful at point of care in a clinical setting when a patient may be seen at only a single visit. New criteria for RA remission, endorsed jointly by the ACR and European League Against Rheumatism, were unveiled at the 2010 ACR Annual Scientific Meeting, with formal publication pending. Unfortunately, no publications yet exist comparing the remission cutoffs for each measurement tool to these new criteria. It is important to recognize that for all of the disease activity measurement tools discussed, it is possible for a patient to meet low disease activity or remission cutoffs and still exhibit residual active disease.

PATIENT (PTGA) AND PROVIDER (PRGA) GLOBAL ASSESSMENT OF DISEASE ACTIVITY

Description

Purpose. The PtGA and PrGA are simple patient-completed or provider-completed visual analog scales (VAS), respectively, measuring the overall way rheumatoid arthritis (RA) affects the patient at a point in time.

Content. The PtGA includes a statement such as "Considering all of the ways your arthritis has affected you, how do you feel your arthritis is today?" with the best anchor and lowest score on the left side and the worst anchor and highest score on the right side (anchors: very well to very poor). The PrGA includes a statement such as "What is your assessment of the patient's current disease activity?" with the best anchor and lowest score on the left side and the worst anchor and highest score on the right side (anchors: none to extremely active).

Developer/contact information. E. C. Huskisson, Department of Rheumatology, St. Bartholomew's Hospital Medical College, Charterhouse Square, London ECIM 6BQ, UK (1977).

Versions. The PtGA and PrGA are classically anchored, unnumbered 10-cm horizontal lines; however, each may be administered as vertical VAS. VAS may be anchored at the ends or open and may have periodic tick marks at specified intervals. Alternatively, a Likert-style VAS may be used. A VAS consisting of 21 circles at 0.5-mm intervals has been shown to be similar to a classic 10-cm line (7).

Number of items. The PrGA and PtGA are each composed of 1 item. There are no subscales.

Populations. The VAS was initially designed to be used in measurement of self-assessed pain in RA (8) and has since been used extensively to evaluate overall disease activity. VAS may be used to measure any 1-dimensional aspect of health.

Practical Application

Method of administration. Patient self-assessment of overall disease activity for PtGA. For PrGA, clinical assessment of overall disease activity performed by the provider.

Administrative burden. None.

Respondent burden. PtGA: patient, ~10 seconds. Provider, ~10 seconds if using a ruler; use of a VAS consisting of 21 circles may be 5 seconds faster than a VAS requiring use of a ruler (7). Laboratory, not applicable. PrGA: same as PtGA, excluding patient time to complete.

Equipment needed. A ruler may be required based on the VAS format chosen.

Availability/cost. There is no cost to use either the PtGA or PrGA.

Scoring. The PtGA and PrGA each range from 0–100 mm, but are often reported from 0–10 cm.

Score interpretation. The level of disease activity increases with higher scores.

Method of scoring. Using a ruler, measure in mm from the left border of the VAS to the point where the patient marked their response on the line. VAS consisting of circles may be scored by visual inspection without use of a ruler. There is no training required to interpret the scores.

Norms available. Proposed definitions of low disease activity are ≤2.0 (scale 0–10) for the PtGA and ≤1.5 (scale 0–10) for the PrGA (9).

Psychometric Information

Reliability. Test–retest reliability for the PtGA based on 122 patients tested on an anchored horizontal 10-cm scale with vertical markers every 10 mm demonstrated an interclass correlation coefficient (ICC) of 0.70 (10). Test–retest reliability for the PrGA based on 166 patient encounters on a horizontal 10-cm scale without vertical markers or anchors demonstrated an ICC for test–retest reliability of 0.96 (10). A study of 22 patients demonstrated a κ of 0.58 and an ICC of 0.44 for the PtGA, while the PrGA demonstrated a κ of 0.79 and an ICC of 0.48 (11). For the PtGA. a study of 24 patients demonstrated an ICC of 0.75 (12).

Validity. *Content.* Both the PtGA and PrGA have historically been used to measure disease in RA. The PtGA and PrGA are both American College of Rheumatology (ACR) core set measures for improvement in RA disease activity. The PtGA is incorporated into many composite indices measuring RA disease activity, including the ACR core data set, Disease Activity Score (DAS) and Disease Activity Score with 28-joint counts (DAS28), Clinical Disease Activity Index (CDAI), and Patient-Based Disease Activity Score without erythrocyte sedimentation rate (ESR); the PrGA is also incorporated into composite indices, including the Simplified Disease Activity Index and CDAI. Most patients evaluate their global assessment of RA disease activity higher than providers (13,14). Interpretation of PrGA and ESR has been shown to demonstrate the least amount of variance among providers as compared to the remaining items in the ACR core set (13).

Concurrent. Both the PtGA and PrGA have been shown to correlate similarly with radiographic scores as compared to other ACR core set measures, with higher values associated with poorer outcomes; however, association with all ACR core set measures was found to be nonsignificant ($P = 0.26$) (13). The level of agreement between DAS scores and PrGA was 49%, and between PrGA and ESR was only 17% (15). Significant correlation between initial and longitudinal analysis between PtGA and PrGA has been shown (16).

Convergent. In clinical trials, the PtGA and PrGA have demonstrated similar change in response to therapy (17,18). In the measurement of pain, vertical and horizontal VAS have been compared with excellent correlation (0.99) between the 2 scales; however, scores from horizontal scales tend to be slightly lower than those from vertical scales (mean ± SD 10.85 ± 0.63 versus 11.05 ± 0.65) (19). Paper- and computer-based versions of the PtGA are highly correlated (ICC 0.91) (20). In 1 study of 24 patients, the 95% limits of agreement for the smallest detectable difference in PtGA ranged between −41 and 32, suggesting poor reliability as compared to multi-item measures of physical function (12).

Construct. Duration of morning stiffness does not correlate with PtGA (21) and when compared to the DAS28, the PtGA was 41% more likely to be concordant than discordant (15). Education may affect patient ratings of disease activity (14), comorbid disease increases the PtGA (14), and support groups may decrease PtGA (22). In a study of 24 patients, the PrGA correlated with the PtGA and pain scores (R = 0.2–0.7), with the Health Assessment Questionnaire (HAQ; R = 0.3–0.7), and with ESR (R = 0.2–0.4) (23). In clinical trials, the PtGA and PrGA have demonstrated a similar change in response to therapy (17,18) and initial PrGA has been shown to correlate with the HAQ 24 months later; however, significant correlation with radiologic outcomes has not been demonstrated (24).

Responsiveness to change. The smallest detectable difference for the PtGA is large as compared to multi-item measures of physical and psychological function and radiologic measures (12), suggesting poor responsiveness to change. However, PrGA and PtGA, pain scores, and the HAQ have been shown to be more sensitive to change than laboratory measures (23). The PrGA correlates with changes in pain scores, morning stiffness, and ACR functional score, but not with physician-derived swollen joint counts (16). One study found that a worsening in PtGA corresponds to a median DAS28 increase of 0.55 (16.5%) and an improvement in PtGA corresponds to a median DAS28 reduction of −0.82 (16.0%; $P < 0.001$) (25).

Critical Appraisal of Overall Value to the Rheumatology Community

The PtGA and PrGA are both components of the ACR core set of measures used to assess RA disease activity and are included in many composite indices. The PtGA and PrGA are practical for use in the clinic as little to no training of staff or patients is required and each may be quickly incorporated into the busy clinical setting. Use of a VAS consisting of 21 circles may be slightly faster than a VAS requiring use of a ruler (7) and overcomes the requirement to carry a ruler, which may be cumbersome. The most important concern threatening the validity and reproducibility of both measures is the lack of uniformity regarding the wording of patient and provider instructions and of anchors. As the PtGA and PrGA are overall measures of patient well-being, the effects of comorbid illness (26) and fixed damage, in addition to current RA disease activity, may influence the score. The PtGA may be more a measure of quality of life than of disease activity as it is decreased after participation in support groups (22) and is influenced by patient education (14). Although the PtGA is a component of many composite measures, when used alone it lacks face validity as no provider-derived data or laboratory parameters are included. Additionally, the PtGA has been shown to correlate poorly with changes in the DAS28 (15,27) and agreement between the PtGA and PrGA is low ($\kappa = 0.17$, $P < 0.001$), indicating that patient perceptions of disease activity may be incongruent with those of providers (28). The PrGA is a component of many composite measures and by itself may encompass all that is known by a provider about a patient's condition, including assessment of joint swelling and pain and interpretation of laboratory values.

DISEASE ACTIVITY SCORE (DAS) AND DISEASE ACTIVITY SCORE WITH 28-JOINT COUNTS (DAS28)

Description

Purpose. The DAS and DAS28 combine single measures into an overall continuous measure of rheumatoid arthritis (RA) disease activity. While both the DAS and DAS28 are feasible to use in RA disease activity monitoring, the shorter DAS28 is more feasible for regular clinical use.

Content. The original DAS includes the number of painful joints calculated by the Ritchie Articular Index (RAI), a 44–swollen joint count (44SJC), erythrocyte sedimentation rate (ESR), and a patient global assessment of disease activity (PtGA) or general health (GH) on a visual analog scale (VAS). The DAS28 includes a 28–swollen joint count (28SJC), 28–tender joint count (28TJC), ESR, and a PtGA or GH assessment on a VAS.

Developer/contact information. P. L. C. M. van Riel, et al, Department of Rheumatology, University Hospital Nijmegen, PO Box 9101, 6500 HB Nijmegen, The Netherlands. E-mail: P.vanriel@reuma.umcn.nl (DAS: 1990, DAS28: 1995).

Versions. The DAS28 is analogous to the DAS; however, the DAS28 includes simplified 28-joint counts. The DAS can be used with or without GH assessment and a PtGA may be substituted for GH. A formula to transform the DAS28 into DAS values is available. The DAS and DAS28 may also be calculated using C-reactive protein (CRP) level instead of Westergren ESR. A total of 8 versions are available.

Number of items. The DAS and DAS28 each have 4 items, or 3 when the GH assessment is omitted. There are no subscales.

Populations. The target population is patients with RA. It is not formally validated for other rheumatic disorders.

Practical Application

Method of administration. Clinical assessment of joint counts and patient assessment of disease activity, combined with laboratory evaluation of ESR.

Administrative burden. Training is necessary for reliable assessment of joint counts.

Time to complete. Patient, ~10 seconds; provider, ~5–8 minutes for DAS and 3–5 minutes for DAS28; laboratory, 1 hour waiting time for ESR, waiting time for CRP level varies by laboratory.

Equipment needed. A ruler may be required based on the VAS format chosen. A laboratory is needed for ESR or CRP determination. A programmed calculator or computer is needed.

Availability/cost. The DAS and DAS28 may be used free of charge and are available online at http://www.das-score.nl.

Scoring. The RAI ranges from 0–78. The 44SJC ranges from 0–44. The 28SJC and 28TJC each range from 0–28. The GH/PtGA VAS ranges from 0–100. The ESR generally ranges from 0–100 mm/hour. Normal CRP level varies by laboratory and is expressed in mg/liter, with use of a CRP test with a lower detection level of 1.0 mg/liter advised. The ranges of the DAS and DAS28 are 0–10 and 0–9.4, respectively.

Scores interpretation. For all DAS versions, the level of disease activity can be interpreted as remission (DAS <1.6), low ($1.6 \leq \text{DAS} < 2.4$), moderate ($2.4 \leq \text{DAS} \leq 3.7$), or high (DAS >3.7) (29). For all DAS28 versions, the level of disease activity can be interpreted as remission (DAS28 <2.6), low ($2.6 \leq \text{DAS28} < 3.2$), moderate ($3.2 \leq \text{DAS28} \leq 5.1$), or high (DAS28 >5.1). Alternative cutoff values for the DAS28 have been proposed that include a more stringent remission of ≤ 2.4 (30).

A change of 1.2 (2 times the measurement error) in either the DAS or DAS28 is considered a significant change (29). The European League Against Rheumatism (EULAR) response criteria classify patients as good, moderate, or nonresponders, using both the change in DAS and level of DAS reached. Good response is defined as improvement in DAS >1.2 and followup DAS ≤ 2.4; nonresponders have improvement in DAS ≤ 0.6 or improvement >0.6 but ≤ 1.2 and followup DAS >3.7; all others are classified as moderate responders (29).

Method of scoring. The following formulas are used: 1) DAS-4 (4 variables) $= 0.54 \times \surd(\text{RAI}) + 0.065 \times (\text{44SJC}) + 0.33 \times \ln(\text{ESR}) + 0.0072 \times \text{GH}$; 2) DAS-3 (3 variables) $= 0.54 \times \surd(\text{RAI}) + 0.065 \times (\text{44SJC}) + 0.33 \times \ln(\text{ESR}) + 0.22$; 3) DAS-4(CRP) $= 0.54 \times \surd(\text{RAI}) + 0.065 \times (\text{44SJC}) + 0.17 \times \ln(\text{CRP} + 1) + 0.0072 \times \text{GH} + 0.45$; 4) DAS-3(CRP) $= 0.54 \times \surd(\text{RAI}) + 0.065 \times (\text{44SJC}) + 0.17 \times \ln(\text{CRP} + 1) + 0.65$; 5) DAS28(4) $= 0.56 \times \surd(\text{28TJC}) + 0.28 \times \surd(\text{28SJC}) + 0.70 \times \ln(\text{ESR}) + 0.014 \times \text{GH}$; 6) DAS28(3) $= [0.56 \times \surd(\text{28TJC}) + 0.28 \times \surd(\text{28SJC}) + 0.70 \times \ln(\text{ESR})] \times 1.08 + 0.16$; 7) DAS28-CRP(4) $= 0.56 \times \surd(\text{28TJC}) + 0.28 \times \surd(\text{28SJC}) + 0.36 \times \ln(\text{CRP} + 1) + 0.014 \times \text{GH} + 0.96$; and 8) DAS28-CRP(3) $= [0.56 \times \surd(\text{28TJC}) + 0.28 \times \surd(\text{28SJC}) + 0.36 \times \ln(\text{CRP} + 1)] \times 1.10 + 1.15$.

The DAS and DAS28 are not directly comparable; however, the following formula may be used to transform scores: $\text{DAS28} = (1.072 \times \text{DAS}) + 0.938$. There is no training required to interpret the scores.

Norms available. Reference values are available and are useful for the interpretation of the DAS from individual patients (see above).

Psychometric Information

Reliability. *Test–retest.* Test–retest reliability for the DAS ranges from 0.8–0.89 (29,31). Test–retest reliability for the DAS with 3 variables was 0.89 (31). When repeated testing 1 week apart was performed, the DAS28 demonstrated a correlation of 0.85 and an interclass correlation coefficient of 0.85 ($P < 0.01$) (32).

Reproducibility. The DAS was designed based on statistical analysis of factors governing clinical judgment in usual rheumatologic care in a cohort of early arthritis patients (33); subsequently, principal component analysis has found no influence of disease duration on the structure of the DAS and DAS28 scores (34).

Validity. *Content.* The DAS and DAS28 include variables from the American College of Rheumatology (ACR) core set of measures used to assess outcomes in RA. No measures of disability or joint damage are included. Due to inclusion of ESR, which is influenced by age, the DAS and DAS28 may underestimate remission in the elderly (35).

Criterion. Physician judgment of disease activity level was used as an external standard for DAS and DAS28 development (33,34). The DAS and DAS28 have been shown to discriminate active from inactive RA disease activity (36), with predictive abilities of 0.93 for the DAS (37) and 0.88 for the DAS28 (38). DAS28 and DAS 28-CRP scores have been shown to discriminate between ACR response categories (39), with good agreement between DAS28-CRP and EULAR response rates at 6 months ($\kappa = 0.80–0.82$) (40,41). DAS28 scores also agree with Clinical Disease Activity Index (CDAI) values ($\kappa = 0.70$) (39). Despite nonsignificance in classification accuracy (42), the frequency of classifying patients in DAS28 remission is higher than for the Simplified Disease Activity Index (SDAI), CDAI (43,44), and DAS (45), and is variable compared to modified ACR criteria (46,47). Misclassification of remission has been shown high as 11.2% when comparing the DAS28-CRP to the DAS28 (48).

Convergent. The DAS and DAS28 are highly correlated ($R = 0.97$) (34), as are DAS28 and DAS28-CRP values ($R = 0.95$); however, correlation between the DAS28 and DAS28-CRP is relatively weak at lower values (48), with DAS28-CRP scores generally lower than DAS28 scores (40). The mean of the differences between the DAS28 ESR- and CRP-based scores is 0.34 (95% limits of agreement −0.58, 1.26) (40). The DAS28 has been shown to explain 9.8% of the total variance in activity restriction (49) and significantly contributes to patient-perceived work disability risk independently from Health Assessment Questionnaire (HAQ) scores (50); however, the DAS28 demonstrates variable levels of correlation with HAQ scores ($R = 0.32–0.68$, $P < 0.05$), which may be related to disease duration (39,51,52). DAS (<1.6) and DAS28 (<2.6) remission may be discordant in up to 15% of patients with 96% of discordant pairs meeting DAS28 but not DAS remission and 4% meeting DAS but not DAS28 remission, suggesting the DAS28 may potentially overestimate remission due to residual disease activity in joints not included in 28-joint counts (45). Approximately 70% of patients classified as being in remission by the DAS28 (DAS28 <2.4) have no swollen joints, while 85% of patients classified as being in remission by the SDAI have no swollen joints (4). The DAS28 correlates highly with provider (R = >0.94) and to a lesser degree with patient assessment of disease activity (R = >0.60) (53). The DAS and DAS28 are variably correlated with power Doppler ultrasound measurement of joint inflammation (54,55).

Predictive. The DAS and DAS28 do not correlate strongly with joint damage at a single point in time (56,57). There is a linear correlation between the DAS28 and radiographic progression (R = 0.58, $P < 0.0001$) (39) with patients who spend >50% of their time in DAS28 remission having less radiographic progression than those not in remission at least 50% of the time (4). EULAR good responders have demonstrated less radiographic progression at up to 18 months ($P = 0.0001$) (58) with EULAR and ACR response criteria, demonstrating equivalent radiographic progression (59). A study evaluating progression of elbow arthritis demonstrated that a cutoff of 3.15 for mean DAS28-CRP(3) at 0–2 years of disease diagnosis predicted elbow deterioration at 10 years (60).

Construct. Increases in the DAS and DAS28 have been shown to correlate with worsening function as measured by the HAQ and Short Form 36 physical functioning scale (51,56,61). The DAS28 and DAS28-CRP have both been shown to have similar mean changes in HAQ scores (−0.86 and −0.80 for good responders, −0.46 and −0.41 for moderate responders, and 0.05 and 0.05 for nonresponders, respectively) (41).

Responsiveness to change. In a trial comparing sulfasalazine with a combination of methotrexate, sulfasalazine, and prednisolone, all composite indices were more responsive than single ACR core set measures. The ACR 20% response criteria were more responsive; the standardized response mean (SRM; mean change divided by SD of change) of the DAS was approximately half as large (2). The SRM for the DAS28 has been shown to be slightly higher than that of the DAS (38).

Critical Appraisal of Overall Value to the Rheumatology Community

Both the DAS and DAS28 have been extensively validated, are endorsed by the ACR and EULAR for RA disease activity measurement in clinical trials (62), and are often considered the "gold standard" by which to measure RA disease activity, with the majority of newer disease activity monitoring tools based on, or compared to, the DAS28. Notably, remission classification for the DAS28 is less conservative than for several other measurement tools, including the DAS, SDAI, and CDAI (40,43,45). The DAS and DAS28 were designed based on statistical analysis of factors governing clinical judgment in usual rheumatologic care in a cohort of early arthritis patients (33,34) with responsiveness of the composite measure repeatedly demonstrated (59). Provider-derived measurement of RA disease activity combined with a laboratory value and a small contribution from patient-derived data gives good face validity. Disadvantages of the DAS and DAS28 in clinical practice are the need for a blood sample, time needed for providers to perform joint counts, complicated mathematical calculation of the composite score, and potential confusion from choosing among the multiple formulas available. Importantly, treating to a target DAS level <2.4 (DAS28 <3.6) has been shown to improve RA outcomes (63,64). Neither the RAI nor the 28TJC and 28SJC can be considered superior (65) and reduced and expanded joint counts have shown similar sensitivity (66), although concern remains that 28-joint counts may miss important disease activity present in uncounted joints (45). While clinicians tend to give swollen joint counts more importance when making treatment decisions (67), the DAS28 formula weighs tender joints more heavily than swollen joints. Additionally, ESR has been shown to contribute 15% of the information in the DAS28 (30) and remission may be underestimated in high ESR states. Conversely, in low ESR states it is possible to be in DAS or DAS28 remission and still have large numbers of swollen joints (68,69). Caution should be employed when using the DAS28-CRP as a substitute for the DAS28, as it produces lower scores (40,41,48) and may underestimate the level of active disease. Limited validation of the DAS28-3 has been performed; however, proposed remission values have been shown to correlate highly with disease activity (R = 0.95, $P < 0.0001$) (70), making the DAS28-3 of use when PtGA data are lacking.

SIMPLIFIED DISEASE ACTIVITY INDEX (SDAI)

Description

Purpose. The SDAI combines single measures into an overall continuous measure of rheumatoid arthritis (RA) disease activity. It is feasible to use for monitoring of RA disease activity in daily clinical practice.

Content. The SDAI includes a 28–swollen joint count (28SJC), 28–tender joint count (28TJC), patient global assessment of disease activity (PtGA) on a 10-cm visual analog scale (VAS), provider global assessment of disease activity (PrGA) on a 10-cm VAS, and C-reactive protein (CRP) level in mg/dl.

Developer/contact information. Josef Smolen, MD, et al, Department of Rheumatology, Medical University of Vienna, Waehringer Guertel 18-20, A-1090 Vienna, Austria. E-mail: Josef.smolen@wienkav.at (2003).

Versions. The SDAI is analogous to the Clinical Disease Activity Index (CDAI); however, the SDAI includes laboratory measurement of CRP level.

Number of items. The SDAI has 5 items. There are no subscales.

Populations. The target population is patients with RA. The SDAI is not validated for other rheumatic disorders.

Practical Application

Method of administration. Clinical assessment of joint counts and PrGA combined with self-administered patient assessment of disease activity.

Administrative burden. Training is necessary for reliable assessment of joint counts.

Time to complete. Patient, ~10 seconds; provider, ~2 minutes; laboratory, waiting time for CRP level varies by laboratory.

Equipment needed. A ruler may be required based on the VAS format chosen. A laboratory is needed for CRP level determination. For calculation of the SDAI, a calculator may be used; however, it is not required.

Availability/cost. The SDAI may be used free of charge; no specialized form is needed.

Scoring. The 28SJC and 28TJC each range from 0–28. The PtGA and PrGA each range from 0–10. Normal CRP level varies by laboratory and is expressed in mg/dl. The lower range of the SDAI is 0, with the upper end resting on the upper limit of CRP level, often defined as 10 mg/dl, leading to an upper limit of 86.

Score interpretation. The level of disease activity can be interpreted as remission (SDAI ≤ 3.3), low ($3.3 <$ SDAI ≤ 11), moderate ($11 <$ SDAI ≤ 26), or high (SDAI >26) (30). The smallest detectable difference for the SDAI is 8.26 (32). An SDAI change of 16 corresponds to a change of 1.2 for the Disease Activity Score with 28-joint counts (DAS28) and an SDAI change of 20.7 corresponds to a 0.22 change in Health Assessment Questionnaire (HAQ) score, both indicating clinically significant change (29,71,72).

Method of scoring. The SDAI is calculated by adding the 5 items together: SDAI = 28SJC + 28TJC + PrGA + PtGA + CRP. There is no training required to interpret the scores.

Norms available. Reference values are available, and are useful for the interpretation of the disease activity scores from individual patients (see above).

Psychometric Information

Reliability. *Test–retest.* When repeated testing 1 week apart was performed, the SDAI demonstrated a correlation of 0.87 and an interclass correlation coefficient of 0.88 ($P <$ 0.01) (32). The components of the SDAI are individually accepted as reliable in RA assessment.

Reproducibility. The SDAI was originally validated with additional patients within the original study (71) and later validated in multiple additional data sets (30). Reproducibility of the SDAI was evaluated in a cross-sectional inception cohort of newly diagnosed RA patients seen every 3 months for 1 year (n = 91) and an observational routine care cohort (n = 279) with significant correlations at the $P < 0.001$ level as compared to the DAS28, CDAI, and HAQ (30).

Validity. *Content.* The SDAI includes variables from the American College of Rheumatology (ACR) core set of measures used to assess outcomes in RA. No measures of disability or joint damage are included. The SDAI was evaluated using a survey of patient profiles among 21 rheumatologists that showed excellent agreement with the physician's ratings of the patients' disease activity (71).

Criterion. The SDAI demonstrates linear correlation with the DAS28 (R = 0.80–0.92, $P < 0.0001$) and is highly related to patient-reported pain (R = 0.66, $P < 0.001$) (73). Major improvement in the SDAI as defined by a decrease of at least 22 at 12 months corresponds to a mean increase of total Sharp score of 1.1, identical to Sharp score progression in good Disease Activity Score (DAS) responders. Larsen scores confirm results of Sharp scores (71). The SDAI is 90% sensitive and 86% specific for prediction of clinicians' decision to change disease-modifying antirheumatic drug therapy and surpasses predictive ability of the DAS28 due to weighting of swollen joint scores (74). Approximately 85% of patients classified as being in remission by the SDAI have no swollen joints; 70% of patients classified as being in remission by the DAS28 have no swollen joints (4). SDAI scores are variably correlated with joint inflammation by power Doppler ultrasound (54,55).

Convergent. Median SDAI scores (11.6, range 0.07–46.60) are slightly higher ($P < 0.001$) than median CDAI scores (10.7, range 0–42.10) and demonstrate sex differential with median SDAI scores of 12.2 (range 0.07–46.60; $P < 0.001$) in women and 8.0 (range 0.10–35.20; $P < 0.001$) in men (73).

Predictive. One study showed that over 3 years, patients who spend the majority of time in SDAI remission do not progress radiographically as compared to those who spent $\leq 50\%$ time in remission (4). The SDAI was sensitive in discriminating between different ACR response categories and HAQ change scores ($P < 0.0001$) (39).

Construct. From baseline to SDAI remission, HAQ score improvement differs by degree of joint damage with ~25% of patients with moderate to severe joint damage having <50% improvement in the HAQ as compared to patients with little joint damage having >80% improvement in the HAQ (75).

Responsiveness to change. As compared to the DAS28, HAQ, and ACR 20% response criteria, the SDAI demonstrates a fairly consistent and proportional change with changes in the SDAI increasing slightly with time (71). Change in the SDAI exhibits a linear relationship with change in the HAQ (R = 0.56–0.57, $P < 0.0001$) and the modified HAQ (R = 0.48, $P < 0.0001$) (71).

Critical Appraisal of Overall Value to the Rheumatology Community

The SDAI was developed to simplify complicated calculations and overcome the shortcomings of the DAS. The SDAI constitutes a simple numerical addition of individual measures on their original scale, overcoming problems of transformations and weighting used in other composite indices. The SDAI is endorsed by the ACR and European League Against Rheumatism (EULAR) for RA disease activity measurement in clinical trials and by EULAR for patient monitoring (6,62). There is no need to use a calculator or computer for calculations. The inclusion of both patient- and provider-derived data as well as a laboratory marker of inflammation gives face validity to the SDAI. The use of CRP in mg/dl, rather than erythrocyte sedimentation rate, is advantageous as it does not overweigh the laboratory variable in the final score. Pain assessment was not included in the SDAI as it is also reflected in the patient global estimate of disease activity. The HAQ disability index was not included in the SDAI as it is not routinely done in clinical practice by many providers (71). Additionally, loss of function as measured by the HAQ or other tools may be irreversible, with some considering addition of a functional measure in a composite index of disease activity confusing despite noting the importance of functional information in patient care (4). Although sex appears to have a small effect on the SDAI score (73), different disease activity cutoffs for male and female patients are not established.

CLINICAL DISEASE ACTIVITY INDEX (CDAI)

Description

Purpose. The CDAI combines single measures into an overall continuous measure of rheumatoid arthritis (RA) disease activity. It is feasible to use the CDAI for monitoring of RA disease activity in daily clinical practice, and lack of a laboratory value may facilitate more routine evaluation of RA disease activity.

Content. The CDAI includes a 28–swollen joint count (28SJC), 28–tender joint count (28TJC), patient global assessment of disease activity (PtGA) on a 10-cm visual analog scale (VAS), and provider global assessment of disease activity (PrGA) on a 10-cm VAS.

Developer/contact information. Daniel Aletaha, MD, et al, Department of Rheumatology, Medical University of Vienna, Waehringer Guertel 18-20, A-1090 Vienna, Austria. E-mail: aletaha@mail.nih.gov (2005).

Versions. The CDAI is analogous to the Simplified Disease Activity Index (SDAI); however, the CDAI excludes laboratory measurement of C-reactive protein (CRP) level.

Number of items. The CDAI has 4 items. There are no subscales.

Populations. The target population is patients with RA. It is not validated for other rheumatic disorders.

Practical Application

Method of administration. Clinical assessment of joint counts and PrGA combined with self-administered patient assessment of disease activity.

Administrative burden. Training is necessary for reliable assessment of joint counts.

Time to complete. Patient, ~10 seconds; provider, <2 minutes; laboratory, not applicable.

Equipment needed. A ruler may be required based on the VAS format chosen. For calculation of the CDAI, a calculator may be used; however, it is not required.

Availability/cost. The CDAI is available free of charge; no specialized form is needed.

Scoring. The 28SJC and 28TJC each range from 0–28. The PtGA and PrGA each range from 0–10. The range of the CDAI is 0–76.

Score interpretation. The level of disease activity can be interpreted as remission (CDAI ≤2.8), low (2.8 < CDAI ≤ 10), moderate (10 < CDAI ≤ 22), or high (CDAI >22) (30), although others have reported differing cutoffs (76). The smallest detectable difference for the CDAI is 8.05 (32).

Method of scoring. The CDAI is calculated by adding the 4 items together: CDAI = 28SJC + 28TJC + PrGA + PtGA. There is no training required to interpret the scores.

Norms available. Reference values are available, and are useful for the interpretation of the disease activity scores from individual patients (see above).

Psychometric Information

Reliability. *Test–retest.* When repeated testing 1 week apart was performed, the CDAI demonstrated a correlation of 0.89 and an interclass correlation coefficient of 0.89 ($P <$ 0.01) (32). The components of the CDAI are individually accepted as reliable in RA assessment.

Reproducibility. Reproducibility of the CDAI was evaluated in a cross-sectional inception cohort of RA patients seen every 3 months for 1 year (n = 91) and observational routine care cohorts (n = 279) with significant correlations ($P < 0.001$) as compared to the Disease Activity Score with 28-joint counts (DAS28), CDAI, and Health Assessment Questionnaire (HAQ) (30). The CDAI was sensitive in discriminating between different American College of Rheumatology (ACR) response categories and HAQ change scores ($P < 0.0001$) (39).

Validity. *Content.* The CDAI includes variables from the ACR core set of measures used to assess outcomes in RA. No measures of disability or joint damage are included. The CDAI has been shown to correlate closely with the SDAI, with only 5% of the variance in the SDAI explained by CRP level (39). The CDAI is highly related to patient-reported pain (R = 0.67, $P < 0.001$) (73).

Criterion. In validation studies, the CDAI showed a correlation coefficient as compared to the DAS28 of 0.89 ($P <$ 0.001) (39). CDAI scores have been shown to correlate with ultrasound-derived swollen joint scores (R = 0.6, $P <$ 0.006) (55).

Convergent. Median CDAI scores (10.7, range 0–42.10) are slightly lower ($P < 0.001$) than median SDAI scores (11.6, range 0.07–46.60) and demonstrate sex differential with median CDAI scores of 11.3 (range 0.00–42.10; $P <$ 0.001) in women and 7.1 (range 0.00–32.00; $P < 0.001$) in men (73).

Predictive. CDAI scores correlate with radiographic change as measured by Larsen scores (R = 0.59, $P <$ 0.0001) (39).

Responsiveness to change. The CDAI demonstrates a linear relationship with the HAQ/modified HAQ (R = 0.47–0.56, $P < 0.0001$) (71) and demonstrates the ability to discriminate degrees of ACR response (39).

Critical Appraisal of Overall Value to the Rheumatology Community

Based on the SDAI, the CDAI was developed as a simple calculation of disease activity for use in the clinic at point of care. The CDAI constitutes a simple numerical addition of individual measures on their original scale, overcoming problems of transformations and weighting used in other composite indices. The CDAI is endorsed by the ACR and European League Against Rheumatism (EULAR) for RA disease activity measurement in clinical trials and by

EULAR for patient monitoring (6,62). There is no need to use a calculator or computer for calculations. Like the SDAI, no functional or quality of life index was included in the CDAI. Like the SDAI, one drawback of the CDAI is not all physicians perform detailed joint examinations when assessing RA disease activity. Additionally, the CDAI does not include any laboratory measurement and therefore all variables are easily available at point of care in the clinical setting, which can in turn produce more consistency in timing and completeness of disease measurement. While exclusion of laboratory measurements decreases face and content validity of the composite measure, similar correlations to the DAS28 and HAQ have been demonstrated as well as discrimination between ACR response categories with and without the addition of CRP level to the remaining variables (39). Additionally, although acute-phase reactants correlate with long-term outcomes in RA, they have not been shown to contribute sufficient information in other composite scores to change judgment of disease activity (39). Although sex appears to have a small effect on the CDAI score (39), different disease activity cutoffs for male and female patients are not established.

PATIENT ACTIVITY SCORE (PAS) AND PATIENT ACTIVITY SCORE-II (PASII)

Description

Purpose. The PAS and PASII combine single measures into an overall continuous measure of rheumatoid arthritis (RA) disease activity. It is feasible to use the PAS/PASII for monitoring of RA disease activity in daily clinical practice as the lack of a laboratory value, or provider-derived data may facilitate more routine evaluation of RA disease activity when time constraints are considered.

Content. The PAS and PASII contain only patient-derived data and include a patient assessment of pain on a 10-cm visual analog scale (VAS), patient global assessment of disease activity (PtGA) on a 10-cm VAS, and the Health Assessment Questionnaire (HAQ) for the PAS or the Health Assessment Questionnaire-II (HAQII) for the PASII.

Developer/contact information. Frederick Wolfe, National Data Bank for Rheumatic Diseases, Arthritis Research Center Foundation, 1035 North Emporia, Suite 288, Wichita, KS 67214. E-mail: fwolfe@arthritis-research.org (2005).

Versions. The PAS is analogous to the PASII and Routine Assessment of Patient Index Data 3 (RAPID3); however, the PASII substitutes the HAQII for the HAQ and RAPID3 substitutes the Multidimensional Health Assessment Questionnaire (MDHAQ) for the HAQ. Spanish versions of the PAS and PASII are available.

Number of items. The PAS and PASII have 3 items each. There are no subscales.

Populations. The target population is patients with RA. The PAS has not been validated for other rheumatic disorders. The PASII has also been used in other rheumatic diseases, including ankylosing spondylitis, psoriatic arthritis, gout, osteoarthritis, fibromyalgia, systemic lupus erythematosus, systemic sclerosis, polymyalgia rheumatica, inflammatory myositis, Sjögren's syndrome, Behçet's disease, and other inflammatory and noninflammatory rheumatic diseases (77).

Practical Application

Method of administration. Patient-administered questionnaire.

Administrative burden. None.

Time to complete. PAS: patient, <3.5 minutes; provider, <1 minute; laboratory, not applicable. PASII: patient, <1.5 minutes; provider, <30 seconds; laboratory, not applicable.

Equipment needed. A ruler may be required based on the VAS format chosen. For calculation of the PAS and PASII, a calculator may be used; however, it is not required.

Availability/cost. The PAS and PASII may be used free of charge and components are available online at http://www.arthritis-research.org/research/pas.

Scoring. The HAQ and HAQII range from 0–3. The patient assessment of pain and PtGA VAS both range from 0–10. The range of both the PAS and PASII is 0–10.

Score interpretation. For the PAS, categories of disease activity may be interpreted as remission (PAS ≤1.25) and low (PAS ≤1.75) (78). For the PASII, categories of disease activity may be defined as remission (PASII ≤1.25) and low (PASII ≤2.2) (43). Categories of self-reported disease severity for the PAS and PASII may be defined as remission = ≤0.25, low = ≤3.7, moderate = <8.0, and high = ≥8.0. (Michaud K, et al: unpublished observations).

Method of scoring. The PAS is calculated by multiplying the HAQ by 3.33 and then dividing the sum of the VAS pain, PtGA, and HAQ by 3. For the PASII, the HAQ is replaced by the HAQII. There is no training required to interpret the scores.

Norms available. Specific reference values are available (see above); however, disease activity cutoffs for the PAS and PASII have been variably reported.

Psychometric Information

Reliability. Test–retest reliability studies have not been done for the PAS or PASII.

Validity. *Content.* The PAS and PASII contain patient-derived variables from the American College of Rheumatology (ACR) core set of measures used to assess outcomes in RA. No provider-derived data are included in the composite scores.

Criterion. The κ value for remission between the PASII and Disease Activity Score with 28-joint counts (DAS28) is 0.25, for the PASII and provider global assessment of disease activity is 0.29, and for the PASII and Clinical Disease Activity Index is 0.40 (43). The PAS has not been evaluated against the Disease Activity Score (DAS), DAS28, or ACR response criteria. An increase of 1 point in the PASII had an odds ratio of 1.19 (95% confidence interval 1.07–1.33) for predicting active from inactive disease independent of specific rheumatic disease diagnosis (77).

Concurrent. The PAS and PASII are equally associated with the Short Form 36 (SF-36) composite scale and the EuroQol (Kendall's τ-a = 0.59). The SF-36 physical and mental component subscales and the Rheumatoid Arthritis Disease Activity Index are more strongly associated with the PAS/PASII as compared to the EuroQol. As compared to the HAQ and HAQII, the PAS and PASII satisfactorily predict mortality, performing only marginally less well (79).

Responsiveness to change. Studies to evaluate for responsiveness to change have not been done for the PAS or PASII. Both the PAS and PASII are comprised of ACR core set measures that have documented responsiveness to change; however, pain scores have been shown to be relatively stable over time in established RA (80).

Critical Appraisal of Overall Value to the Rheumatology Community

The PAS and PASII are simple to use composite scores that use patient-derived data without provider effort, which makes the feasibility of the measure very good for use in daily clinical care. While the PAS and PASII are composed of the 3 patient-derived ACR core set data elements, the measures have not been compared directly to measures such as the ACR response criteria or the DAS in clinical trials and may not be suitable for use in that setting. As the components of the PAS are included in the majority of observational and clinical studies, it has been suggested that the PAS could be used to compare data between such studies (79). Although the HAQII is not widely used at this time, it possesses superior psychometric properties to the MHAQ and MDHAQ, and a formula exists to convert from the HAQII to the HAQ (81,82). Other than the choice of HAQ version, the PAS is identical to both the PASII and RAPID3. An advantage of the PASII as compared to the PAS is use of the shorter HAQII, which is therefore easier for patients to complete. Time to complete the PASII is identical to RAPID3; however, the HAQII is impacted less by skipped questions and has a lesser floor effect as compared to the MDHAQ (82).

ROUTINE ASSESSMENT OF PATIENT INDEX DATA (RAPID)

Description

Purpose. The RAPID2 and RAPID3 are brief self-administered questionnaires of disease symptoms for patients with rheumatoid arthritis (RA). RAPID4 patient joint count (RAPID4PTJC) adds a self-reported joint count to RAPID3. RAPID4 provider joint count (RAPID4MDJC) and RAPID5 combine self-administered questionnaires of disease symptoms with a provider-derived joint count or provider global assessment of disease activity (PrGA), respectively.

Content. The RAPID scores include combinations of the following 2–5 items of the following 6 items: the Multidimensional Health Assessment Questionnaire (MDHAQ), a pain visual analog scale (VAS), patient global assessment of disease activity (PtGA) on a 10-cm VAS, PrGA on a 10-cm VAS, swollen joint count (SJC), and tender joint count (TJC).

Developer/contact information. Theodore Pincus, MD, Director of Outcomes Research, Division of Rheumatology, Department of Medicine, New York University Hospital for Joint Diseases, New York University Medical Center, 301 East 17th Street, Room 1608, New York, New York 10003. Email: tedpincus@gmail.com (1999).

Versions. There are 5 RAPID score versions: RAPID2 includes the PtGA and the PrGA; RAPID3 includes the MDHAQ, a pain VAS, and the PtGA; RAPID4TJC includes the MDHAQ, a pain VAS, PtGA, and the Rheumatoid Arthritis Disease Activity Index (RADAI) self-reported TJC; RAPID4MDJC includes the MDHAQ, a pain VAS, PtGA, and provider-derived 28TJC and 26SJC (excludes shoulders); and RAPID5 includes the MDHAQ, a pain VAS, PtGA, PrGA, and the RADAI self-reported TJC (83,84). A Spanish version is available.

Number of items. The RAPID scores have 2–5 items, corresponding to the number after RAPID (i.e., RAPID3 has 3 items). There are no subscales.

Populations. The target population is patients with RA. RAPID3 has been used in systemic lupus erythematosus, spondylarthropathies, vasculitis, psoriatic arthritis, gout, scleroderma, osteoarthritis, fibromyalgia, familial Mediterranean fever, and Behçet's disease (85).

Practical Application

Method of administration. For RAPID score versions requiring a joint count, either provider joint counts or patient self-reported RADAI joint counts may be used to calculate RAPID scores (84).

Administrative burden. Training is necessary for reliable assessment of joint counts.

Time to complete. RAPID2: patient, ~20 seconds; provider, <20 seconds; laboratory, not applicable (N/A). RAPID3: patient, ~1.5 minutes; provider, <30 seconds; laboratory, N/A. RAPID4PTJC: patient, 5–10 minutes; provider, <1 minute; laboratory: N/A. RAPID4MDJC: patient, <1.5 minutes; provider, ~2 minutes; laboratory, N/A. RAPID5: patient, 5–10 minutes if patient joint count and <5 minutes if provider joint count; provider, <1 minute if patient joint count and <2.5 minutes if provider joint count; laboratory, N/A (86–88).

Equipment needed. A ruler may be required based on the VAS format chosen. For calculation of RAPID scores a scoring template or calculator may be used, but is not required.

Availability/cost. The RAPID scores are all available free of charge. The RAPID3 is available online at http://mdhaq.org/Public/Questionnaires.aspx.

Scoring. The RADAI self-report joint count ranges from 0–48. Provider tender joint scores range from 0–28. Provider swollen joint scores range from 0–26. The MDHAQ ranges from 0–3. Pain, PtGA, and PrGA are scored from 0–10, each on a separate 10-cm horizontal VAS, which may be substituted for a 21-circle VAS. The range of each of the RAPID scores is 0–10. Scores can be calculated by hand, by calculator, or by use of a scoring template.

Score interpretation. Recommended cutoffs for disease activity are not available for RAPID2 scores. RAPID3, RAPID4PTJC, RAPID4MDJC, and RAPID5 scores range from 0–1.0 = remission, 1.1–2.0 = low, 2.1–4.0 = moderate, and 4.1–10 = high (89). Alternatively, if the RAPID3 is scored on a 0–30 scale, the following cutoffs for disease activity may be used: remission = 0–3.0, low = 3.1–6.0, moderate = 6.1–12.0, and high = 12.1–30 (86). The smallest detectable difference for RAPID3 is 1.48 (32).

Method of scoring. All raw scores may be converted to a range of 0–10. The MDHAQ score (0–3) is multiplied by 3.33. Pain, PtGA, and PrGA are scored from 0–10, each on a separate 10-cm VAS. Tender and swollen joint scores are converted to a 0–10 scale based on simple division (i.e., for 28-joint count, divide by 2.8). After the component scores are standardization to 0–10 scales, the individual items in the desired composite are added together and then divided by the number of items in the composite to give an adjusted final score (i.e., RAPID3 raw score of 0–30 is divided by 3 to give the final adjusted score ranging 0–10) (84). There is no training required to interpret the scores.

Norms available. See above.

Psychometric Information

Reliability. When repeated testing 1 week apart was performed, RAPID3 demonstrated a correlation of 0.88 and an interclass correlation coefficient of 0.90 ($P < 0.01$) (32). Test–retest reliability has not been evaluated for the other RAPID scores.

Validity. *Content.* The RAPID scores are all composed of American College of Rheumatology (ACR) core set measures used to assess the efficacy of disease-modifying antirheumatic drugs. None of the RAPID scores includes acute-phase reactant values.

Criterion. The RAPID3, RAPID4MDJC, RAPID4PTJC, and RAPID5 scores demonstrate Spearman's correlation coefficients as compared to the Disease Activity Score with 28-joint counts (DAS28) between 0.69 and 0.73, with the RAPID4MDJC demonstrating the highest value. Spearman's correlation coefficients for each of the RAPID scores (excluding RAPID2) as compared to the Clinical Disease Activity Index (CDAI) ranged from 0.74–0.83, with the RAPID4MDJC again demonstrating the highest value. Correlation coefficients between each of the RAPID scores (excluding RAPID2) were between 0.98 and 0.99. An additional study demonstrated Spearman's correlation coefficients between 0.36 and 0.61 for RAPID3 as compared to the DAS28 and between 0.54 and 0.77 as compared to the CDAI (86).

Concurrent. When comparing RAPID3 to ACR 20% improvement criteria and the DAS in a randomized controlled trial, the κ ranged between 0.62 and 0.64, indicating similar sensitivity in identification of active treatment from placebo (90). In another study, κ values ranged from 0.22–0.37 for CDAI versus RAPID3 and from 0.12–0.20 for DAS28 versus CDAI, indicating a fair agreement of RAPID3 to both of the indices (86).

Construct. Addition of TJC or SJCs and/or physician estimate of global status does not add to the capacity of RAPID3 to distinguish active from control treatments, and all RAPID scores perform similarly in this function (84).

Responsiveness to change. Studies to evaluate for responsiveness to change have not been done for the RAPID scores. Each of the RAPID scores is comprised of combinations of ACR core set measures that have well-documented responsiveness to change; however, pain scores have been shown to be relatively stable over time in established RA (80).

Critical Appraisal of Overall Value to the Rheumatology Community

The RAPID scores are feasible in the clinical setting and are based on ACR criteria. The flexibility to choose which of the RAPID scores used is advantageous. The RAPID3 score is the most frequently used and best-validated measure among the various RAPID scores. RAPID4 scores have been presented in 2 different forms with use of TJCs or both TJCs and SJCs and as such, standardization is lacking. Additionally, RAPID4 and RAPID5 scores have been reported in the literature using either provider- or patient-derived joint counts ranging from 28–78 joints. This lack of standardization makes these scores more difficult to compare between users. Furthermore, both provider and patient joint counts are poorly reproducible (91–93) and patient-reported TJCs correlate only moderately with physician joint counts, with SJCs demonstrating lower levels of correlation (94). More recently, Pincus et al have advocated using the RADAI self-reported joint count (95), which would partially eliminate issues of variability among joint assessment. While RAPID2 scores are the easiest to perform in a busy clinical setting, they also provide the least information as compared to the other RAPID scores. Patient-derived measures without acute-phase reactant values have been shown to be reliable and sensitive (96); however, they may be influenced by patient education level (97). While the use of the MDHAQ is shorter and easier to administer than the original HAQ, it has a greater floor effect and missing 1 or more items has more impact on the score than the newer HAQ-II (81,82), which is a component of the otherwise identical Patient Activity Score-II. Additionally, despite concerns that patient questionnaires may reflect irreversible joint damage, RAPID scores have demonstrated similar efficiency to joint counts in differentiating active from inactive disease (86).

RHEUMATOID ARTHRITIS DISEASE ACTIVITY INDEX (RADAI) AND RHEUMATOID ARTHRITIS DISEASE ACTIVITY INDEX-5 (RADAI-5)

Description

Purpose. The RADAI and RADAI-5 are patient-assessed measures for disease activity in rheumatoid arthritis (RA). Both the RADAI and RADAI-5 may complement or replace the physician's assessment of disease activity in health services or epidemiologic research, and may be used for patient management.

Content. The RADAI and RADAI-5 are both 5-item questionnaires. The items ask about global disease activity in the last 6 months, current disease activity with respect to joint tenderness and swelling, arthritis pain, and duration of morning stiffness. The RADAI also includes tender joints rated on a joint list and the RADAI-5 asks about general health.

Developer/contact information. RADAI: Gerold Stucki, Department of Physical Medicine and Rehabilitation, University Hospital Munich, Marchionistrasse 15, D-81377 Munich, Germany. E-mail: gerold.stucki@phys.med.uin-muenchen.de (1995). RADAI-5: Burkhard F. Leeb, Karl Landsteiner Institute for Clinical Rheumatology, Landstrasse 18, A-2000 Stockerau, Austria. E-mail: burkhard.leeb@stockerau.lknoe.at (2008).

Versions. The RADAI-5 is analogous to the RADAI with replacement of joint counts by the patient's general health assessment and slight change of question format. The RADAI is also available in German, French, Italian, and Dutch. The RADAI-5 is also available in German.

Number of items. The RADAI and RADAI-5 each contain 5 items. There are no subscales.

Populations. The target population is patients with RA. The RADAI has been evaluated for use in patients with undifferentiated peripheral inflammatory arthritis (98). The RADAI-5 may be useful in other forms of inflammatory arthritis.

Practical Application

Method of administration. Both the RADAI and RADAI-5 are self-administered questionnaires designed to be self-explanatory.

Administrative burden. None.

Time to complete. RADAI: patient, <5 minutes; provider, <30 seconds; laboratory, not applicable (N/A). RADAI-5: patient, <1 minute; provider, <30 seconds; laboratory, N/A (99).

Equipment needed. A calculator or computer may be used to calculate the RADAI and RADAI-5.

Availability/cost. The RADAI and RADAI-5 are available free of charge.

Scoring. *RADAI.* Items 1–3 are scored on numerical rating scales from 0 = "not at all" to 10 = "extremely/very severe." Item 4 on morning stiffness is scored on an ordered category scale from 0 = "none," 1 = "<30 minutes," 2 = "30 minutes to 1 hour," 3 = "1–2 hours," 4 = "2–4 hours," 5 = ">4 hours," and 6 = "all day" and transformed to a 0–10 range. Item 5, the joint list, is calculated as the sum of 16 joints or joint groups scored from 0 indicating "no pain" to 3 indicating "severe pain," with the 0–48 score transformed to the 0–10 range.

RADAI-5. All items are scored on a 0–10 ordered category scale. For items 1 and 2, 0 = "completely inactive" and 10 = "extremely active." For item 3, 0 = "no pain" and 10 = "unbearable pain." For item 4, 0 = "very good" health status and 10 = "very bad" health status. For item 5, 0 = "no stiffness" and 10 = "stiffness the whole day."

Both the RADAI and the RADAI-5 are calculated as the mean of the nonmissing items and range from 0–10.

Score interpretation. Both RADAI and RADAI-5 scores of 0 mean no RA disease activity, with higher scores indicating higher levels of disease activity. Within individual patients, a change in RADAI >1.4 can be interpreted as clinically meaningful (38,100).

Method of scoring. The RADAI can be scored with a calculator or computer. The RADAI-5 can be scored by hand or with a calculator. There is no training required to interpret scores.

Norms available. A RADAI score ≤2.0 has been identified as low or inactive disease (nonflare) (38); however, additional disease categories have not been established. Disease activity categories for the RADAI-5 have been proposed: 0.0–1.4 = a remission-like state, 1.6–3.0 = mild disease activity, 3.2–5.4 = moderate disease activity, and 5.6–10.0 = high disease activity (99).

Psychometric Information

Reliability. *Test–retest reliability.* Test–retest correlation for the RADAI performed within 1 week was high, with an interclass correlation coefficient of 0.92 (100). Test–retest reliability has not been proven for the RADAI-5.

Internal consistency. For the RADAI, Cronbach's α of 0.87 and 0.91 have been reported (101,102). For the RADAI-5, a Cronbach's α of 0.92 has been reported. Factor analysis by principal component analysis reveals both the RADAI and RADAI-5 to be 1-dimensional instruments with all items contributing significantly to the final scores (101,103).

Validity. *Content.* The RADAI and RADAI-5 items score the most apparent symptoms of RA with the exception of joint scores. No physician-derived measures of disease activity or laboratory parameters are included in either measurement tool.

Criterion. The RADAI is significantly correlated ($P < 0.01$) at low to moderate degrees with tender joint counts (TJCs; 0.44–0.55) and swollen joint counts (SJCs; 0.39–0.54), and variably correlated with acute-phase reactants (101,102). The RADAI-5 is significantly correlated ($P < 0.001$) with TJCs (0.60) and SJCs (0.60); however, it was not correlated significantly with erythrocyte sedimentation rate or C-reactive protein (CRP) level (103). The κ for the relationship between the RADAI-5 and the original RADAI was excellent at 0.88, with fair agreement for the RADAI-5 as compared to the Disease Activity Score with 28-joint counts (DAS28; 0.29), DAS28-CRP (0.34), Simplified Disease Activity Index (SDAI; 0.39), and Clinical Disease Activity Index (CDAI; 0.37) (103).

Predictive. The distribution of the RADAI is skewed toward normal, with up to 14% of patients scoring <1, preventing numerical improvement despite potential clinical improvement. Spearman's ρ demonstrated a linear relationship between change in RADAI-5 scores compared to change in DAS28 (0.59, $P < 0.001$) and change in CDAI (0.57, $P < 0.001$). Low to moderate κ correlations for the relationship between change in the RADAI-5 and change in the DAS28 (0.30) and change in the CDAI (0.54) have been demonstrated (103).

Convergent. Spearman's rank correlations between the RADAI and RADAI-5 demonstrate near perfect correlation

(0.995, $P < 0.001$). The RADAI is significantly correlated ($P < 0.01$) to a moderate degree with the Health Assessment Questionnaire (0.55–0.57) and DAS28 (0.53) (101,102). The RADAI and RADAI-5 demonstrated correlations with the provider global assessment of disease activity (PrGA) of 0.54–0.59 and 0.60 ($P < 0.01$), respectively, compared to a correlation of 0.72 ($P < 0.01$) between PrGA and DAS28 (101–103). Statistically significant correlations between the RADAI-5 and DAS28, DAS28-CRP, SDAI, and CDAI were also found, with Spearman's ρ ranging from 0.64–0.74 ($P < 0.001$) (103).

Responsiveness to change. The RADAI has been shown to detect flare of RA with equal sensitivity to the DAS28 (predictive ability 0.88) (38) and to discriminate European League Against Rheumatism good responders from moderate and nonresponders with predictive ability of 0.78 (100). For the RADAI-5, despite significant changes for individual patients, no significant changes were seen at the group level between 2 assessments done within 3 months (103).

Critical Appraisal of Overall Value to the Rheumatology Community

The RADAI and RADAI-5 are short, easy to understand, patient-administered tools that produce a single index of disease activity, designed for feasibility in a busy clinical setting. The RADAI was designed to provide a continuous score useful in clinical and epidemiologic research and is an aggregate of selected items from the 5 categories of the Rapid Assessment of Disease Activity in Rheumatology (RADAR) (101). The RADAR questionnaire is not discussed in this report as it does not provide a continuous score. The lack of physician, laboratory, and joint swelling assessments in the RADAI and lack of specificity to RA produced by the use of joint groups (i.e., fingers instead of metacarpophalangeal and proximal interphalangeal joints) detracts from the face validity of the measure. Additionally the RADAI has shown lower correlation with PrGA than DAS28 scores (102). The RADAI-5 improves upon the original RADAI in terms of simplicity of calculation, and the omission of joint counts makes assessment faster. Additionally, the disease activity categories of the RADAI-5 demonstrate a more normal distribution than DAS28 and CDAI scores (25). Despite lacking assessment of joint counts or physician-derived components, the RADAI-5 has been shown to correlate with these measures on an individual basis as well as with other composite measures that contain such items (103,104). Relationships between both the RADAI and RADAI-5 and acute-phase reactant values have been weak or absent, diminishing the face validity of the composite measures (101–104). The RADAI lacks well-defined disease states, both the RADAI and RADAI-5 have not been validated outside of white cohorts (103), and the RADAI-5 has not been used in the research setting with sensitivity to change and relationships to disease outcomes, such as bony erosions or prediction of disability undocumented.

CHRONIC ARTHRITIS SYSTEMIC INDEX (CASI)

Description

Purpose. The CASI combines single measures into an overall continuous measure of rheumatoid arthritis (RA) disease activity. It is feasible to use for monitoring of RA disease activity in daily clinical practice.

Content. The CASI includes the Ritchie Articular Index (RAI), patient assessment of pain visual analog scale (VAS), Health Assessment Questionnaire (HAQ), and erythrocyte sedimentation rate (ESR).

Developer/contact information. G. F. Ferraccioli, MD, Division and Chair of Rheumatology, DPMSC, University of Udine, 33100, Udine, Italy. E-mail: gianfranco.ferraccioli@dpmsc.uniud.it (1993).

Versions. None.

Number of items. The CASI has 4 items. There are no subscales.

Populations. The target population is patients with RA. It is not validated for other rheumatic disorders.

Practical Application

Method of administration. Clinical assessment of joint counts and drawing a blood sample for ESR combined with self-administered patient assessments of disease activity and pain.

Administrative burden. Training is necessary for reliable assessment of joint counts.

Time to complete. Patient, <3.5 minutes; provider, ~3 minutes; laboratory, 1 hour waiting time for ESR.

Equipment needed. A ruler may be required based on the VAS format chosen. A laboratory is needed for ESR determination. For calculation of the CASI, a calculator or computer may be used.

Availability/cost. The CASI may be used free of charge; no specialized form is needed.

Scoring. The RAI ranges from 0–78. The HAQ ranges from 0–3. The pain VAS ranges from 0–100. ESR generally ranges from 0–100. The range of the CASI is 0–74.

Score interpretation. The level of disease activity can be interpreted as CASI remission ≤24.65; this corresponds to a Disease Activity Score (DAS) ≤3.32. Additional categories of disease activity have not been established (37).

Method of scoring. The following formula is used: $\text{CASI} = 13 \times \text{HAQ} + 0.21 \times \text{ESR} + 0.08 \times \text{pain VAS} + 0.07692 \times \text{RAI}$. There is no training required to interpret the scores.

Norms available. Other than remission, disease activity categories have not been established.

Psychometric Information

Reliability. *Test–retest.* The test–retest reliability of the CASI was tested after a period of 1 week with a test–retest correlation of 0.86 (105).

Internal consistency. Cronbach's α of 0.81 has been reported (105).

Reproducibility. The CASI was validated in a cross-sectional study inclusive of additional patients seen within the same clinic (37).

Validity. *Content.* The CASI includes variables from the American College of Rheumatology core set of measures used to assess outcomes in RA.

Concurrent. One study over 6 months demonstrated a statistically significant decrease in CASI (39.9 to 24.7; $P < 0.01$), while at the same time demonstrating similar decreases in the HAQ (1.69 to 1.01; $P < 0.05$), ESR (55.2 to 34; $P < 0.01$), RAI (21.5 to 10.9; $P < 0.01$), and pain VAS (66.6 to 34.5; $P < 0.01$) (105). The CASI correlates with the DAS ($R = 0.545$, $P < 0.01$) (106) and with several indices of RA disease activity not currently in wide usage: the Thompson Articular Index (0.56, $P < 0.01$), the Lansbury Index (0.88, $P < 0.01$), the Stoke Index (0.44, $P < 0.01$), and the Mallya and Mace Index (0.53, $P < 0.01$) (105).

Construct. The 4 factors composing the CASI contributed 74% of the common variance in a study of 124 patients: HAQ (39%), pain VAS (6%), ESR (8%), and RAI (21%). The CASI demonstrates a negative correlation with grip strength (0.44, $P < 0.01$) and positive correlations with number of swollen joints (0.39, $P < 0.01$), morning stiffness (0.42, $P < 0.01$), and the Zung Depression Inventory (0.52, $P < 0.01$) (105).

Responsiveness to change. The CASI demonstrates similar sensitivity and specificity for remission as compared to the DAS (37). Over 6 months the CASI shows similar change over time as compared to ESR (106), C-reactive protein level, fibrinogen levels, morning stiffness, proximal interphalangeal joint synovitis score, pain levels tested by the Present Pain Index and the McGill Pain Questionnaire, the Performance and Activity Scale, and the Lansbury Index ($P < 0.01$), while changes in the Arthritis Impact Measurement Scales were nonsignificant in this population (105).

Critical Appraisal of Overall Value to the Rheumatology Community

The CASI was designed using factorial analysis of 29 available variables with the intent to design an RA measure of both disease activity and severity for use by practicing rheumatologists. Validation studies were mainly performed on measures no longer in general use. The inclusion of both patient- and provider-derived data as well as a laboratory marker of inflammation gives face validity to the CASI. The CASI correlates moderately with the DAS; however, the suggested cutoff for remission corresponds to a DAS much higher than the currently accepted remission level of 1.6. Patient global assessment of disease activity was not included in lieu of patient pain assessment as they were thought to measure the same thing, a claim supported by the high correlation between the 2 items (0.90, $P < 0.01$) (37). The calculation for the CASI requires use of a calculator or computer, which may make the measure difficult to use at point of care. The use of the RAI increases the time required to perform joint counts as compared to standard joint counts as grading of tenderness is required. Additionally, inclusion of ESR may make point of care use difficult in clinics that do not have laboratory values available at the time a patient is examined. The inclusion of the original HAQ increases the time required of the patient to complete the measure as compared to measures using shorter versions of the HAQ or alternative quality of life measures, and may not add sufficient additional information to justify the increase in time spent. Further studies are needed to define and validate categories of disease activity.

PATIENT-BASED DISEASE ACTIVITY SCORE WITH ESR (PDAS1) AND PATIENT-BASED DISEASE ACTIVITY SCORE WITHOUT ESR (PDAS2)

Description

Purpose. The PDAS1 and PDAS2 combine single measures into an overall continuous measure of rheumatoid arthritis (RA) disease activity. The PDAS1 and PDAS2 are feasible to use for monitoring of RA disease activity in daily clinical practice.

Content. The PDAS1 includes a patient-assessed 50–tender joint count (50TJC), patient global assessment of disease activity (PtGA), the modified Health Assessment Questionnaire (MHAQ), and erythrocyte sedimentation rate (ESR). The PDAS2 includes a patient-assessed 28–swollen joint count (28SJC), PtGA, the MHAQ, and an early morning stiffness (EMS) score.

Developer/contact information. Ernest H. Choy, MS, ROCP, Sir Alfred Baring Garrod Clinical Trials Unit, Academic Department of Rheumatology, King's College London, Weston Education Centre, Cutcombe Road, London SE5 9RJ, UK. E-mail: ernest.choy@kcl.ac.uk (2008).

Versions. The PDAS2 is similar in design to the PDAS1; however, the PDAS1 includes laboratory measurement of ESR, includes a patient-assessed 50TJC instead of a patient-assessed 28SJC, and excludes the EMS score.

Number of items. The PDAS1 and PDAS2 each have 4 items. There are no subscales.

Populations. The target population is patients with RA. It is not validated for other rheumatic disorders.

Practical Application

Method of administration. Self-administered TJCs (PDAS1) or SJCs (PDAS2) using a mannequin displaying individual joints, patient self-assessments of disease activity, and physical function combined with laboratory evaluation of ESR (PDAS1) or self-administered EMS scale (PDAS2).

Administrative burden. Training is necessary for reliable assessment of joint counts.

Time to complete. PDAS1: patient, 5–10 minutes (107); provider, <1 minute; laboratory: 1 hour waiting time for ESR. PDAS2: patient, 5–10 minutes (107); provider, <1 minute; laboratory, not applicable.

Equipment needed. A ruler may be required based on the visual analog scale (VAS) format chosen. A laboratory is needed for ESR determination. For calculation of the PDAS1 and PDAS2, a calculator or computer is needed.

Availability/cost. The PDAS1 and PDAS2 are available free of charge.

Scoring. The 50TJC ranges from 0–50. The self-administered 28SJC ranges from 0–28. The PtGA ranges from 0–10. The MHAQ ranges from 0–3. The ESR generally ranges from 0–100. The EMS score ranges from 0–5. The range of the PDAS1 is 0–7.6. The range of the PDAS2 is 0–2.6.

Score interpretation. The PDAS1 and PDAS2 produce continuous scores. Ranges for disease activity have not been established.

Method of scoring. The following formulas are used: PDAS1 = 0.119 × (PtGA) + 0.842 × ln(ESR + 2) + 0.432 × ln(patient 50TJC + 2) + 0.271 × (MHAQ), and PDAS2 = 0.021 × (PtGA) + 0.483 × (MHAQ) + 0.033 × (self-administered 28SJC) + 0.002 × (EMS). There is no training required to interpret the scores.

Norms available. Reference values have not been established.

Psychometric Information

Reliability. During development of the tool, test–retest reliability was assessed for each item within 24 hours of initial assessment and was graded as excellent, with interclass correlation coefficients ranging from 0.76–0.88 for both the PDAS1 and PDAS2 (107).

Validity. *Content.* The PDAS1 includes 3 of the 7, and the PDAS2 includes 2 of the 7, American College of Rheumatology (ACR) core set measures used to assess outcomes in RA. While self-assessed joint counts are not included in the ACR core set of measures, self-assessed joint counts have been shown to variably correlate with joint counts performed by physicians (R = 0.41–0.78), but are not equivalent, with patients rating joint tenderness higher than physicians (60,61,108–110). Patient training has been shown to improve correlation with physician counts (R = 0.94) (110). In the development study, assessor 28TJCs demonstrated a correlation of 0.53 with the PDAS1 and a correlation of 0.76 with the Disease Activity Score with 28-joint counts (DAS28) (107).

Criterion. The PDAS2 is less sensitive than both the PDAS1 and DAS28 for detecting low disease activity (107).

Construct. The PDAS1 and PDAS2 showed convergent validity with moderate correlations with other measures of RA disease activity (PDAS1: DAS28 = 0.89, Clinical Disease Activity Index [CDAI] = 0.69, provider global assessment of disease activity [PrGA] VAS = 0.68, VAS = 0.72; PDAS2: DAS28 = 0.76, CDAI = 0.73, PrGA VAS = 0.67, VAS pain = 0.83). Both the PDAS1 and PDAS2 demonstrate divergent validity, with correlations between 0.23 and 0.37 for the PDAS1 and between 0.30 and 0.44 for the PDAS2 with the Nottingham Health Profile sleep, social, and emotion scales (107).

Responsiveness to change. The PDAS1 and PDAS2 demonstrate similar sensitivity to change as compared to the DAS28, with a mean ± SD change over 6 months of 1.00 ± 1.30 for the PDAS1 and 0.73 ± 1.10 for the PDAS2 versus 1.20 ± 1.50 for the DAS28 (107).

Critical Appraisal of Overall Value to the Rheumatology Community

The PDAS1 and PDAS2 were designed to provide the information produced by the DAS28 in a simpler fashion in order to facilitate routine use in the clinic. As not all physicians perform detailed joint examinations when assessing RA disease activity, use of patient self-administered joint counts may be a good alternative to physician-administered joint counts (109). The format employed by the PDAS1 and PDAS2 may be superior to patient assessment of joints in list format (94,107). The subjective nature of self-administered TJCs included in the PDAS1 may be easier for providers to accept than self-administered SJCs employed by some measures such as the PDAS2. Although simple patient training has been shown to improve quality of self-administered joint counts (110), patient training may impose additional burden on busy staff. Requirement of a 50-joint count in the PDAS1 is more extensive than most other currently used measures and may be cumbersome for patients, while not providing a significant amount of additional information. Additionally, lack of provider-derived information limits face validity of the measures. The PDAS1 requires laboratory measurement of ESR, which while adding face validity, may not be available at point of care in the clinical setting. The PDAS2 also uses an EMS scale, which is not used in many RA disease activity assessment tools at this time, despite evidence demonstrating that duration of morning stiffness has a moderate ability to differentiate active from inactive RA (111). Another drawback of both the PDAS1 and PDAS2 is that the complicated calculations, including transformation of variables, require use of a calculator. While a valid quality of life measure, and of the choices in HAQ versions, the MHAQ lacks a normal distribution, clusters at the lower end of the scale, and may fail to show numerical improvement in up to 25% of patients despite clinical improvement in function (82,112–115), thus making the composite PDAS1 and PDAS2 less responsive at the lower end of the scale. Additionally, further studies are needed to define and validate categories of disease activity.

MEAN OVERALL INDEX FOR RHEUMATOID ARTHRITIS (MOI-RA)

Description

Purpose. The MOI-RA combines single measures into an overall continuous measure of RA disease activity. It is feasible to use the MOI-RA for monitoring of RA disease activity in daily clinical practice.

Content. The MOI-RA includes the mean of standardized values of tender and swollen joint counts (28-, 42-, or 66/68-joint counts), patient global assessment of disease activity (PtGA) on a 10-cm visual analog scale (VAS), patient assessment of pain on a 10-cm VAS, provider global assessment of disease activity (PrGA) on a 10-cm VAS, Health Assessment Questionnaire (HAQ), and erythrocyte sedimentation rate (ESR; range 0–100).

Developer/contact information. Heidi Makinen, Jyvaskyla Central Hospital, Keskussairaalantie 19, 40620 Jyvaskyla, Finland (2008).

Versions. The MOI-RA can be used with 28-, 42-, or 66/68-joint counts.

Number of items. The MOI-RA has 7 items. There are no subscales.

Populations. The target population is patients with RA. There are no other uses.

Practical Application

Method of administration. Clinical assessment of joint counts and provider assessment of global disease activity, drawing a blood sample for ESR, and patient-administered HAQ, and pain and patient global assessment of disease activity VAS.

Administrative burden. Training is necessary for reliable assessment of joint counts.

Time to complete. Patient, 5–10 minutes; provider, 3–8 minutes; laboratory, 1 hour waiting time for ESR.

Equipment needed. A ruler may be required based on the VAS format chosen. A laboratory is needed for ESR determination. For calculation of the MOI-RA, a calculator or computer is needed.

Availability/cost. None.

Scoring. All components are standardized to range from 0–100. VAS for PtGA, patient assessment of pain, and PrGA each range from 0–100. ESR values >100 are replaced by the value 100. HAQ values ranging from 0–3 are multiplied by 100 (28). The range of the MOI-RA is 0–100.

Score interpretation. Higher values indicate poorer outcomes.

Method of scoring. The mean of the standardized values is calculated producing a range for the MOI-RA between 0 and 100, with higher values indicating poorer outcomes. If values of 1–3 components are missing, standardized values are calculated from the available component values and the mean of the standardized values is recorded. The MOI-RA is most easily scored with use of a calculator. There is no training required to interpret scores.

Norms available. Norms have not been documented for the MOI-RA.

Psychometric Information

Reliability. Test–retest reliability has not been studied for the aggregate measure. The components of the MOI-RA are individually accepted as reliable in RA assessment.

Validity. *Content.* The MOI-RA is a composite measure inclusive of all 7 of the American College of Rheumatology (ACR) core set measures used to assess outcomes in RA chosen by expert consensus to represent valid dimensions of disease activity in RA.

Criterion. The MOI-RA was compared to ACR response criteria with mean change in the MOI-RA from baseline to 6 months similar in patients grouped into those who did not meet ACR 20% response criteria (ACR20), those who met ACR20 but not ACR50 response, those with ACR50 but not ACR remission, and those who met ACR remission criteria. The MOI-RA was also compared to the Disease Activity Score with 28-joint counts (DAS28) with correlations between 0.84 and 0.90 over a period of 6 months (28).

Construct. Changes in the MOI-RA were correlated with changes in the DAS28, which has been shown to correlate with disability over time (61) and contains the HAQ, which has itself demonstrated the ability to distinguish between placebo and treatment groups (116) with changes over time agreeing with and augmenting clinical and laboratory evidence of change (80,117,118).

Predictive. No studies on predictive validity have been done. The MOI-RA has been shown to correlate with the DAS28 and HAQ, which have each been shown to predict RA-related morbidity (119,120).

Responsiveness to change. In the Finnish Rheumatoid Arthritis Combination Therapy (FIN-RACo) study, MOI-RA values decreased by ~65% with a corresponding decrease in the mean DAS28 of 50%. Both the MOI-RA and DAS28 were able to significantly discriminate the 2 treatment arms of the FIN-RACo study (28).

Critical Appraisal of Overall Value to the Rheumatology Community

The MOI-RA was developed based on data for 169 RA patients with complete ACR core data; however, test–retest reliability of the composite measure has not been established. With inclusion of all 7 of the ACR core set measures, the MOI-RA has excellent face validity; however, it has not been widely used to this date, and as such the MOI-RA has not been further validated in additional RA cohorts. Disadvantages of the MOI-RA in clinical practice are the need for a blood sample, the time needed for providers to perform joint counts, and complicated mathematical calculation of the composite score. Additionally, cutoffs for categories of disease activity have not been established.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

ROLE OF THE STUDY SPONSOR

No author had any financial support or other benefits from commercial sources for the work reported on in the manuscript. Abbott Laboratories had no financial interest in this project and had no input in the design or content with all opinions and conclusions expressed herein those of the authors.

REFERENCES

1. Fransen J, Stucki G, van Riel PL. Rheumatoid arthritis measures: Disease Activity Score (DAS), Disease Activity Score-28 (DAS28), Rapid Assessment of Disease Activity in Rheumatology (RADAR), and Rheumatoid Arthritis Disease Activity Index (RADAI). Arthritis Rheum 2003;49 Suppl:S214–24.
2. Verhoeven AC, Boers M, van Der Linden S. Responsiveness of the core set, response criteria, and utilities in early rheumatoid arthritis. Ann Rheum Dis 2000;59:966–74.
3. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al, for the Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.
4. Aletaha D, Smolen JS. Remission of rheumatoid arthritis: should we care about definitions? Clin Exp Rheumatol 2006;24 Suppl:S45–51.
5. Ledingham J, Deighton C, for the British Society for Rheumatology

Standards, Guidelines and Audit Working Group. Update on the British Society for Rheumatology guidelines for prescribing TNFα blockers in adults with rheumatoid arthritis (update of previous guidelines of April 2001). Rheumatology (Oxford) 2005;44:157–63.

6. Smolen JS, Landewe R, Breedveld FC, Dougados M, Emery P, Gaujoux-Viala C, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs. Ann Rheum Dis 2010;69:964–75.
7. Pincus T, Bergman M, Sokka T, Roth J, Swearingen C, Yazici Y. Visual analog scales in formats other than a 10 centimeter horizontal line to assess pain and other clinical data. J Rheumatol 2008;35:1550–8.
8. Scott PJ, Huskisson EC. Measurement of functional capacity with visual analogue scales. Rheumatol Rehabil 1977;16:257–9.
9. Wells GA, Boers M, Shea B, Brooks PM, Simon LS, Strand CV, et al. Minimal disease activity for rheumatoid arthritis: a preliminary definition. J Rheumatol 2005;32:2016–24.
10. Rohekar G, Pope J. Test-retest reliability of patient global assessment and physician global assessment in rheumatoid arthritis. J Rheumatol 2009;36:2178–82.
11. Hernandez-Cruz B, Cardiel MH. Intra-observer reliability of commonly used outcome measures in rheumatoid arthritis. Clin Exp Rheumatol 1998;16:459–62.
12. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol 2001;28:892–903.
13. Aletaha D, Machold KP, Nell VP, Smolen JS. The perception of rheumatoid arthritis core set measures by rheumatologists: results of a survey. Rheumatology (Oxford) 2006;45:1133–9.
14. Nicolau G, Yogui MM, Vallochi TL, Gianini RJ, Laurindo IM, Novaes GS. Sources of discrepancy in patient and physician global assessments of rheumatoid arthritis disease activity. J Rheumatol 2004;31:1293–6.
15. Wolfe F, Michaud K, Pincus T, Furst D, Keystone E. The Disease Activity Score is not suitable as the sole criterion for initiation and evaluation of anti–tumor necrosis factor therapy in the clinic: discordance between assessment measures and limitations in questionnaire use for regulatory purposes. Arthritis Rheum 2005;52:3873–9.
16. Hanly JG, Mosher D, Sutton E, Weerasinghe S, Theriault D. Self-assessment of disease activity by patients with rheumatoid arthritis. J Rheumatol 1996;23:1531–8.
17. Strand V, Cohen S, Crawford B, Smolen JS, Scott DL, for the Leflunomide Investigators Groups. Patient-reported outcomes better discriminate active treatment from placebo in randomized controlled trials in rheumatoid arthritis. Rheumatology (Oxford) 2004;43:640–7.
18. Van Riel PL, Freundlich B, MacPeek D, Pedersen R, Foehl JR, Singh A, et al. Patient-reported health outcomes in a trial of etanercept monotherapy versus combination therapy with etanercept and methotrexate for rheumatoid arthritis: the ADORE trial. Ann Rheum Dis 2008;67:1104–10.
19. Scott J, Huskisson EC. Vertical or horizontal visual analogue scales. Ann Rheum Dis 1979;38:560.
20. Athale N, Sturley A, Skoczen S, Kavanaugh A, Lenert L. A web-compatible instrument for measuring self-reported disease activity in arthritis. J Rheumatol 2004;31:223–8.
21. Yazici Y, Erkan D, Peterson MG, Kagen LJ. Morning stiffness: how common is it and does it correlate with physician and patient global assessment of disease activity? [letter]. J Rheumatol 2001;28:1468–9.
22. Baker PR, Groh JD, Kraag GR, Tugwell P, Wells GA, Boisvert D. Impact of patient with patient interaction on perceived rheumatoid arthritis overall disease status. Scand J Rheumatol 1996;25:207–12.
23. Ward MM. Clinical measures in rheumatoid arthritis: which are most useful in assessing patients? J Rheumatol 1994;21:17–27.
24. Berglin E, Lorentzon R, Nordmark L, Nilsson-Sojka B, Rantapaa Dahlqvist S. Predictors of radiological progression and changes in hand bone density in early rheumatoid arthritis. Rheumatology (Oxford) 2003;42:268–75.
25. Rintelen B, Sautner J, Haindl P, Andel I, Maktari A, Leeb B. Comparison of three rheumatoid arthritis disease activity scores in clinical routine. Scand J Rheumatol 2009;7:1–7.
26. Harrington JT. The uses of disease activity scoring and the physician global assessment of disease activity for managing rheumatoid arthritis in rheumatology practice. J Rheumatol 2009;36:925–9.
27. Leeb BF, Sautner J, Leeb BA, Fassl C, Rintelen B. Lack of agreement between patients and physicians perspectives of rheumatoid arthritis disease activity changes. Scand J Rheumatol 2006;35:441–6.
28. Makinen H, Kautiainen H, Hannonen P, Sokka T. A new disease activity index for rheumatoid arthritis: Mean Overall Index for Rheumatoid Arthritis (MOI-RA). J Rheumatol 2008;35:1522–7.
29. Van Gestel AM, Prevoo ML, van 't Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis: comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism criteria. Arthritis Rheum 1996;39:34–40.
30. Aletaha D, Smolen J. The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. Clin Exp Rheumatol 2005;23 Suppl:S100–8.
31. Fransen J, van Riel PL. The Disease Activity Score and the EULAR response criteria. Clin Exp Rheumatol 2005;23 Suppl:S93–9.
32. Uhlig T, Kvien TK, Pincus T. Test-retest reliability of disease activity core set measures and indices in rheumatoid arthritis. Ann Rheum Dis 2009;68:972–5.
33. Van der Heijde DM, van 't Hof MA, van Riel PL, Theunisse LA, Lubberts EW, van Leeuwen MA, et al. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. Ann Rheum Dis 1990;49:916–20.
34. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight–joint counts: development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44–8.
35. Radovits BJ, Fransen J, van Riel PL, Laan RF. Influence of age and gender on the 28-joint Disease Activity Score (DAS28) in rheumatoid arthritis. Ann Rheum Dis 2008;67:1127–31.
36. Van der Heijde DM, van 't Hof M, van Riel PL, van de Putte LB. Validity of single variables and indices to measure disease activity in rheumatoid arthritis. J Rheumatol 1993;20:538–41.
37. Salaffi F, Peroni M, Ferraccioli GF. Discriminating ability of composite indices for measuring disease activity in rheumatoid arthritis: a comparison of the Chronic Arthritis Systemic Index, Disease Activity Score and Thompson's Articular Index. Rheumatology (Oxford) 2000;39:90–6.
38. Fransen J, Hauselmann H, Michel BA, Caravatti M, Stucki G. Responsiveness of the self-assessed Rheumatoid Arthritis Disease Activity Index to a flare of disease activity. Arthritis Rheum 2001;44:53–60.
39. Aletaha D, Nell VP, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther 2005;7:R796–806.
40. Hensor EM, Emery P, Bingham SJ, Conaghan PG, for the YEAR Consortium. Discrepancies in categorizing rheumatoid arthritis patients by DAS-28(ESR) and DAS-28(CRP): can they be reduced? Rheumatology (Oxford) 2010;49:1521–9.
41. Wells G, Becker JC, Teng J, Dougados M, Schiff M, Smolen J, et al. Validation of the 28-joint Disease Activity Score (DAS28) and European League Against Rheumatism response criteria based on C-reactive protein against disease progression in patients with rheumatoid arthritis, and comparison with the DAS28 based on erythrocyte sedimentation rate. Ann Rheum Dis 2009;68:954–60.
42. Aletaha D. Pooled indices to measure rheumatoid arthritis activity: a good reflection of the physician's mind? Arthritis Res Ther 2006;8:102.
43. Shaver TS, Anderson JD, Weidensaul DN, Shahouri SS, Busch RE, Mikuls TR, et al. The problem of rheumatoid arthritis disease activity and remission in clinical practice. J Rheumatol 2008;35:1015–22.
44. Smolen JS, Aletaha D. What should be our treatment goal in rheumatoid arthritis today? Clin Exp Rheumatol 2006;24 Suppl:S7–13.
45. Landewe R, van der Heijde D, van der Linden S, Boers M. Twenty-eight-joint counts invalidate the DAS28 remission definition owing to the omission of the lower extremity joints: a comparison with the original DAS remission. Ann Rheum Dis 2006;65:637–41.
46. Listing J, Strangfeld A, Rau R, Kekow J, Gromnica-Ihle E, Klopsch T, et al. Clinical and functional remission: even though biologics are superior to conventional DMARDs overall success rates remain low: results from RABBIT, the German biologics register. Arthritis Res Ther 2006;8:R66.
47. Atzeni F, Antivalle M, Pallavicini FB, Caporali R, Bazzani C, Gorla R, et al. Predicting response to anti-TNF treatment in rheumatoid arthritis patients. Autoimmun Rev 2009;8:431–7.
48. Inoue E, Yamanaka H, Hara M, Tomatsu T, Kamatani N. Comparison of Disease Activity Score (DAS)28-erythrocyte sedimentation rate and DAS28 C-reactive protein threshold values. Ann Rheum Dis 2007;66:407–9.
49. Kuhlow H, Fransen J, Ewert T, Stucki G, Forster A, Langenegger T, et al. Factors explaining limitations in activities and restrictions in participation in rheumatoid arthritis. Eur J Phys Rehabil Med 2010;46:169–77.
50. Macedo A, Oakley S, Gullick N, Kirkham B. An examination of work instability, functional impairment, and disease activity in employed patients with rheumatoid arthritis. J Rheumatol 2009;36:225–30.
51. Drossaers-Bakker KW, de Buck M, van Zeben D, Zwinderman AH, Breedveld FC, Hazes JM. Long-term course and outcome of functional

capacity in rheumatoid arthritis: the effect of disease activity and radiologic damage over time. Arthritis Rheum 1999;42:1854–60.

52. Prajs K, Flicinski J, Brzosko I, Przepiera-Bedzak H, Ostanek L, Brzosko M. Quality of life and activity of disease in patients with rheumatoid arthritis. Ann Acad Med Stetin 2006;52 Suppl:39–43. In Polish.
53. Villaverde V, Balsa A, Cantalejo M, Fernandez-Prada M, Madero MR, Munoz-Fernandez S, et al. Activity indices in rheumatoid arthritis. J Rheumatol 2000;27:2576–81.
54. Balsa A, de Miguel E, Castillo C, Peiteado D, Martin-Mola E. Superiority of SDAI over DAS-28 in assessment of remission in rheumatoid arthritis patients using power Doppler ultrasonography as a gold standard. Rheumatology (Oxford) 2010;49:683–90.
55. Kawashiri SY, Kawakami A, Iwamoto N, Fujikawa K, Satoh K, Tamai M, et al. The power Doppler ultrasonography score from 24 synovial sites or 6 simplified synovial sites, including the metacarpophalangeal joints, reflects the clinical disease activity and level of serum biomarkers in patients with rheumatoid arthritis. Rheumatology (Oxford) 2010;50:962–5.
56. Fransen J, Uebelhart D, Stucki G, Langenegger T, Seitz M, Michel BA. The ICIDH-2 as a framework for the assessment of functioning and disability in rheumatoid arthritis. Ann Rheum Dis 2002;61:225–31.
57. Seror R, Tubach F, Baron G, Guillemin F, Ravaud P. Measure of function in rheumatoid arthritis: individualized or classical scales? Ann Rheum Dis 2010;69:97–101.
58. Van Gestel AM, Haagsma CJ, van Riel PL. Validation of rheumatoid arthritis improvement criteria that include simplified joint counts. Arthritis Rheum 1998;41:1845–50.
59. Van Gestel AM, Anderson JJ, van Riel PL, Boers M, Haagsma CJ, Rich B, et al. ACR and EULAR improvement criteria have comparable validity in rheumatoid arthritis trials. J Rheumatol 1999;26:705–11.
60. Abe A, Ishikawa H, Murasawa A, Nakazono K. Disease activity and the course of elbow joint deterioration over 10 years in the patients with early rheumatoid arthritis. Clin Rheumatol 2008;27:867–72.
61. Welsing PM, van Gestel AM, Swinkels HL, Kiemeney LA, van Riel PL. The relationship between disease activity, joint destruction, and functional capacity over the course of rheumatoid arthritis. Arthritis Rheum 2001;44:2009–17.
62. Aletaha D, Landewe R, Karonitsch T, Bathon J, Boers M, Bombardier C, et al. Reporting disease activity in clinical trials of patients with rheumatoid arthritis: EULAR/ACR collaborative recommendations. Ann Rheum Dis 2008;67:1360–4.
63. Goekoop-Ruiterman YP, de Vries-Bouwstra JK, Kerstens PJ, Nielen MM, Vos K, van Schaardenburg D, et al. DAS-driven therapy versus routine care in patients with recent-onset active rheumatoid arthritis. Ann Rheum Dis 2010;69:65–9.
64. Grigor C, Capell H, Stirling A, McMahon AD, Lock P, Vallance R, et al. Effect of a treatment strategy of tight control for rheumatoid arthritis (the TICORA study): a single-blind randomised controlled trial. Lancet 2004;364:263–9.
65. Prevoo ML, van Riel PL, van 't Hof MA, van Rijswijk MH, van Leeuwen MA, Kuper HH, et al. Validity and reliability of joint indices: a longitudinal study in patients with recent onset rheumatoid arthritis. Br J Rheumatol 1993;32:589–94.
66. Fuchs HA, Pincus T. Reduced joint counts in controlled clinical trials in rheumatoid arthritis. Arthritis Rheum 1994;37:470–5.
67. Lindsay K, Ibrahim G, Sokoll K, Tripathi M, Melsom RD, Helliwell PS. The composite DAS score is impractical to use in daily practice: evidence that physicians use the objective component of the DAS in decision making. J Clin Rheumatol 2009;15:223–5.
68. Fransen J, van Riel PL. DAS remission cut points. Clin Exp Rheumatol 2006;24 Suppl:S29–32.
69. Smolen JS, Aletaha D. Interleukin-6 receptor inhibition with tocilizumab and attainment of disease remission in rheumatoid arthritis: the role of acute-phase reactants. Arthritis Rheum 2011;63:43–52.
70. Balsa A, Carmona L, Gonzalez-Alvaro I, Belmonte MA, Tena X, Sanmarti R, et al. Value of Disease Activity Score 28 (DAS28) and DAS28–3 compared to American College of Rheumatology-defined remission in rheumatoid arthritis. J Rheumatol 2004;31:40–6.
71. Smolen JS, Breedveld FC, Schiff MH, Kalden JR, Emery P, Eberl G, et al. A simplified disease activity index for rheumatoid arthritis for use in clinical practice. Rheumatology (Oxford) 2003;42:244–57.
72. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557–60.
73. Rintelen B, Haindl PM, Maktari A, Nothnagl T, Hartl E, Leeb BF. SDAI/CDAI levels in rheumatoid arthritis patients are highly dependent on patient's pain perception and gender. Scand J Rheumatol 2008;37:410–3.
74. Soubrier M, Zerkak D, Gossec L, Ayral X, Roux C, Dougados M. Which variables best predict change in rheumatoid arthritis therapy in daily clinical practice? J Rheumatol 2006;33:1243–6.
75. Aletaha D, Smolen J, Ward MM. Measuring function in rheumatoid arthritis: identifying reversible and irreversible components. Arthritis Rheum 2006;54:2784–92.
76. Arya V, Malaviyab AN, Raja RR. CDAI (Clinical Disease Activity Index) in rheumatoid arthritis: cut-off values for classification into different grades of disease activity. Indian J Rheumatol 2007;2:91–4.
77. Parekh K, Taylor WJ. The Patient Activity Scale-II is a generic indicator of active disease in patients with rheumatic disorders. J Rheumatol 2010;37:1932–4.
78. Wolfe F, Rasker JJ, Boers M, Wells GA, Michaud K. Minimal disease activity, remission, and the long-term outcomes of rheumatoid arthritis. Arthritis Rheum 2007;57:935–42.
79. Wolfe F, Michaud K, Pincus T. A composite disease activity scale for clinical practice, observational studies, and clinical trials: the Patient Activity Scale (PAS/PAS-II). J Rheumatol 2005;32:2410–5.
80. Wolfe F, Hawley DJ, Cathey MA. Clinical and health status measures over time: prognosis and outcome assessment in rheumatoid arthritis. J Rheumatol 1991;18:1290–7.
81. Anderson J, Sayles H, Curtis JR, Wolfe F, Michaud K. Converting Modified Health Assessment Questionnaire (HAQ), Multidimensional HAQ, and HAQII scores into original HAQ scores using models developed with a large cohort of rheumatoid arthritis patients. Arthritis Care Res (Hoboken) 2010;62:1481–8.
82. Wolfe F, Michaud K, Pincus T. Development and validation of the Health Assessment Questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum 2004;50:3296–305.
83. Pincus T. A Multidimensional Health Assessment Questionnaire (MDHAQ) for all patients with rheumatic diseases to complete at all visits in standard clinical care. Bull NYU Hosp Jt Dis 2007;65:150–60.
84. Pincus T, Bergman MJ, Yazici Y, Hines P, Raghupathi K, Maclean R. An index of only patient-reported outcome measures, Routine Assessment of Patient Index Data 3 (RAPID3), in two abatacept clinical trials: similar results to Disease Activity Score (DAS28) and other RAPID indices that include physician-reported measures. Rheumatology (Oxford) 2008;47:345–9.
85. Pincus T, Sokka T. Can a Multi-Dimensional Health Assessment Questionnaire (MDHAQ) and Routine Assessment of Patient Index Data (RAPID) scores be informative in patients with all rheumatic diseases? Best Pract Res Clin Rheumatol 2007;21:733–53.
86. Pincus T, Swearingen CJ, Bergman MJ, Colglazier CL, Kaell AT, Kunath AM. RAPID3 (Routine Assessment of Patient Index Data) on an MDHAQ (Multidimensional Health Assessment Questionnaire): agreement with DAS28 (Disease Activity Score) and CDAI (Clinical Disease Activity Index) activity categories, scored in five versus more than ninety seconds. Arthritis Care Res (Hoboken) 2010;62:181–9.
87. Pincus T, Yazici Y, Bergman M. A practical guide to scoring a Multi-Dimensional Health Assessment Questionnaire (MDHAQ) and Routine Assessment of Patient Index Data (RAPID) scores in 10-20 seconds for use in standard clinical care, without rulers, calculators, websites or computers. Best Pract Res Clin Rheumatol 2007;21:755–87.
88. Yazici Y, Bergman M, Pincus T. Time to score quantitative rheumatoid arthritis measures: 28-joint count, Disease Activity Score, Health Assessment Questionnaire (HAQ), Multidimensional HAQ (MDHAQ), and Routine Assessment of Patient Index Data (RAPID) scores. J Rheumatol 2008;35:603–9.
89. Pincus T, Swearingen CJ, Bergman M, Yazici Y. RAPID3 (Routine Assessment of Patient Index Data 3), a rheumatoid arthritis index without formal joint counts for routine care: proposed severity categories compared to disease activity score and clinical disease activity index categories. J Rheumatol 2008;35:2136–47.
90. Pincus T, Strand V, Koch G, Amara I, Crawford B, Wolfe F, et al. An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the Disease Activity Score (DAS) in a rheumatoid arthritis clinical trial. Arthritis Rheum 2003;48:625–30.
91. Scott DL, Choy EH, Greeves A, Isenberg D, Kassinor D, Rankin E, et al. Standardising joint assessment in rheumatoid arthritis. Clin Rheumatol 1996;15:579–82.
92. Klinkhoff AV, Bellamy N, Bombardier C, Carette S, Chalmers A, Esdaile JM, et al. An experiment in reducing interobserver variability of the examination for joint tenderness. J Rheumatol 1988;15:492–4.
93. Thompson PW, Hart LE, Goldsmith CH, Spector TD, Bell MJ, Ramsden MF. Comparison of four articular indices for use in clinical trials in rheumatoid arthritis: patient, order and observer variation. J Rheumatol 1991;18:661–5.
94. Barton JL. Systematic review and metaanalysis of patient self-report versus trained assessor joint counts in rheumatoid arthritis. J Rheumatol 2009;36:2635–41.
95. Pincus T, Yazici Y, Bergman M, Maclean R, Harrington T. A proposed continuous quality improvement approach to assessment and management of patients with rheumatoid arthritis without formal joint counts, based on quantitative Routine Assessment of Patient Index

Data (RAPID) scores on a Multidimensional Health Assessment Questionnaire (MDHAQ). Best Pract Res Clin Rheumatol 2007;21:789−804.

96. Pincus T, Callahan LF, Brooks RH, Fuchs HA, Olsen NJ, Kaye JJ. Self-report questionnaire scores in rheumatoid arthritis compared with traditional physical, radiographic, and laboratory measures. Ann Intern Med 1989;110:259−66.

97. Callahan LF, Smith WJ, Pincus T. Self-report questionnaires in five rheumatic diseases: comparisons of health status constructs and associations with formal education level. Arthritis Care Res 1989;2:122–31.

98. Castrejon I, Silva-Fernandez L, Bombardier C, Carmona L. Clinical composite measures of disease activity for diagnosis and followup of undifferentiated peripheral inflammatory arthritis: a systematic review. J Rheumatol Suppl 2011;87:48−53.

99. Rintelen B, Haindl PM, Sautner J, Leeb BA, Deutsch C, Leeb BF. The Rheumatoid Arthritis Disease Activity Index-5 in daily use: proposal for disease activity categories. J Rheumatol 2009;36:918−24.

100. Fransen J, Forster A, Uebelhart D, Michel BA. Reliability and responsiveness of the RADAI, a self-assessed rheumatoid arthritis disease activity index [abstract]. Ann Rheum Dis 2001;60 Suppl:345.

101. Stucki G, Liang MH, Stucki S, Bruhlmann P, Michel BA. A self-administered Rheumatoid Arthritis Disease Activity Index (RADAI) for epidemiologic research: psychometric properties and correlation with parameters of disease activity. Arthritis Rheum 1995;38:795−8.

102. Fransen J, Langenegger T, Michel BA, Stucki G. Feasibility and validity of the RADAI, a self-administered rheumatoid arthritis disease activity index. Rheumatology (Oxford) 2000;39:321−7.

103. Leeb BF, Haindl PM, Maktari A, Nothnagl T, Rintelen B. Patient-centered rheumatoid arthritis disease activity assessment by a modified RADAI. J Rheumatol 2008;35:1294−9.

104. Leeb BF, Sautner J, Mai HT, Haindl PM, Deutsch C, Rintelen B. A comparison of patient questionnaires and composite indexes in routine care of rheumatoid arthritis patients. Joint Bone Spine 2009;76: 658−64.

105. Ferraccioli GF, Salaffi F, Troise-Rioda W, Bartoli E. The Chronic Arthritis Systemic Index (CASI). Clin Exp Rheumatol 1994;12:241−7.

106. Ferraccioli G, Bartoli E, Salaffi F, Peroni M. The Chronic Arthritis Systemic Index: a nomogram to assess the activity and severity of chronic arthritis [letter]. Arthritis Rheum 1993;36:1180−1.

107. Choy EH, Khoshaba B, Cooper D, MacGregor A, Scott DL. Development and validation of a patient-based disease activity score in rheumatoid arthritis that can be used in clinical trials and routine practice. Arthritis Rheum 2008;59:192–9.

108. Calvo FA, Calvo A, Berrocal A, Pevez C, Romero F, Vega E, et al. Self-administered joint counts in rheumatoid arthritis: comparison with standard joint counts. J Rheumatol 1999;26:536−9.

109. Figueroa F, Braun-Moscovici Y, Khanna D, Voon E, Gallardo L, Luinstra D, et al. Patient self-administered joint tenderness counts in rheumatoid arthritis are reliable and responsive to changes in disease activity. J Rheumatol 2007;34:54−6.

110. Levy G, Cheetham C, Cheatwood A, Burchette R. Validation of patient-reported joint counts in rheumatoid arthritis and the role of training. J Rheumatol 2007;34:1261–5.

111. Khan NA, Yazici Y, Calvo-Alen J, Dadoniene J, Gossec L, Hansen TM, et al. Reevaluation of the role of duration of morning stiffness in the assessment of rheumatoid arthritis activity. J Rheumatol 2009;36: 2435–42.

112. Stucki G, Stucki S, Bruhlmann P, Michel BA. Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. Ann Rheum Dis 1995;54: 461–5.

113. Wolfe F. Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. J Rheumatol 2001;28:982–9.

114. Martin M, Kosinski M, Bjorner JB, Ware JE Jr, Maclean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res 2007;16:647–60.

115. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly Health Assessment Questionnaire format. Arthritis Rheum 1999;42: 2220−30.

116. Egsmose C, Lund B, Borg G, Pettersson H, Berg E, Brodin U, et al. Patients with rheumatoid arthritis benefit from early 2nd line therapy: 5 year followup of a prospective double blind placebo controlled study. J Rheumatol 1995;22:2208−13.

117. Pincus T, Sokka T. Quantitative measures for assessing rheumatoid arthritis in clinical trials and clinical care. Best Pract Res Clin Rheumatol 2003;17:753−81.

118. Fitzpatrick R, Newman S, Lamb R, Shipley M. A comparison of measures of health status in rheumatoid arthritis. Br J Rheumatol 1989;28:201−6.

119. Wagner E, Ammer K, Kolarz G, Krajnc I, Palkonyai E, Scherak O, et al. Predicting factors for severity of rheumatoid arthritis: a prospective multicenter cohort study of 172 patients over 3 years. Rheumatol Int 2007;27:1041−8.

120. Leigh JP, Fries JF, Parikh N. Severity of disability and duration of disease in rheumatoid arthritis. J Rheumatol 1992;19:1906−11.

121. Wolfe F, Kleinheksel SM, Cathey MA, Hawley DJ, Spitz PW, Fries JF. The clinical value of the Stanford Health Assessment Questionnaire Functional Disability Index in patients with rheumatoid arthritis. J Rheumatol 1988;15:1480−8.

Summary Table for RA Outcome Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden†	Score interpretation	Reliability evidence‡	Validity evidence‡	Ability to detect change‡	Strengths	Cautions
DAS	RA disease activity measurement RAI: 0–78, SJC44: 0–44, ESR: 0–100, PtGA VAS: 0–100	Provider assessment, patient item, lab	Patient: ~10 sec	Provider: 5–8 min Lab: 1 hour waiting time for ESR	Remission: DAS <1.6 Low: 1.6 ≤ DAS < 2.4 Moderate: 2.4 ≤ DAS ≤ 3.7 High: DAS >3.7	Excellent	Excellent	Excellent	Discriminates well between remission and active RA Used extensively in clinical trials Includes acute-phase reactant, patient- and provider-derived data	Requires provider joint counts, acute-phase reactant, calculator/computer for calculation Heavily weighted for ESR Not validated in non-RA disease states
DAS28	RA disease activity measurement TJC28: 0–28, SJC28: 0–28, ESR: 0–100, PtGA VAS: 0–100	Provider assessment, patient item, lab	Patient: ~10 sec	Provider: 3–5 min Lab: 1 hour waiting time for ESR	Remission: DAS28 <2.6 Low: 2.6 ≤ DAS28 < 3.2 Moderate: 3.2 ≤ DAS28 ≤ 5.1 High: DAS28 >5.1	Excellent	Excellent	Excellent	Discriminates well between remission and active RA Used extensively in clinical trials Includes acute-phase reactant, patient- and provider-derived data	Requires provider joint counts, acute-phase reactant, calculator/computer for calculation Heavily weighted for ESR Not validated in non-RA disease states
SDAI	RA disease activity measurement TJC28: 0–28, SJC28: 0–28, PtGA VAS: 0–10, CRP: 0–10	Provider item, patient item, lab	Patient: ~10 sec	Provider: ~2 min Lab: waiting time for CRP varies by lab	Remission: SDAI ≤3.3 Low: 3.3 < SDAI ≤ 11 Moderate: 11 < SDAI ≤ 26 High: SDAI >26	Good, test–retest reliability for composite has not been evaluated	Excellent	Excellent	Simple mathematical scoring without transformation Includes acute-phase reactant, patient- and provider-derived data	Requires provider joint counts, acute-phase reactant Not validated in non-RA disease states
CDAI	RA disease activity measurement TJC28: 0–28, SJC28: 0–28, PtGA VAS: 0–10, PrGA VAS: 0–10	Provider item, patient item	Patient: ~10 sec	Provider: <2 min Lab: N/A	Remission: CDAI ≤2.8 Low: 2.8 < CDAI ≤ 10 Moderate 10 < CDAI ≤ 22 High: CDAI >22	Good, test–retest reliability for composite has not been evaluated	Excellent	Excellent	Simple mathematical scoring without transformation Includes patient- and provider-derived data All variables easily available at point of care	Requires provider joint counts No acute-phase reactant included Not validated in non-RA disease states
PAS	RA disease activity measurement HAQ: 0–3; pain VAS: 0–10, PtGA VAS: 0–10	Patient questionnaire	Patient: <3.5 min	Provider: <1 min Lab: N/A	Remission: ≤0.25 Minimal: ≤3.7 Moderate: <8.0 Severe: ≥8.0	Acceptable, test–retest reliability for composite has not been evaluated	Acceptable	Acceptable for individual components, has not been evaluated for composite	Simple mathematical scoring without transformation Includes patient-derived data only	No provider-derived data Not validated in non-RA disease states May be influenced by patient education level
PASII	RA disease activity measurement HAQII: 0–3, pain VAS: 0–10, PtGA VAS: 0–10	Patient questionnaire	Patient: <1.5 min	Provider: <30 sec Lab: N/A	Remission: ≤0.25 Minimal: ≤3.7 Moderate: <8.0 Severe: ≥8.0	Acceptable, test–retest reliability for composite has not been evaluated	Acceptable	Acceptable for individual components, has not been evaluated for composite	Simple mathematical scoring without transformation Includes patient-derived data only	No provider-derived data Not validated in non-RA disease states May be influenced by patient education level
RAPID2	Disease activity measurement in rheumatic disease PtGA VAS: 0–10, PrGA VAS: 0–10	Provider item, patient item	Patient: 20 sec	Provider: <20 sec Lab: N/A	Specific cutoffs not available, higher scores indicate greater disease activity	Acceptable, test–retest reliability for composite has not been evaluated	Poor	Acceptable for individual component, has not been evaluated for composite	Simple and easy to perform Includes patient- and provider-derived data	No acute-phase reactant included May be influenced by patient education level Least information as compared to the other RAPID scores

(continued)

Summary Table *(Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden†	Score interpretation	Reliability evidence‡	Validity evidence‡	Ability to detect change‡	Strengths	Cautions
RAPID3	Disease activity measurement in rheumatic disease MDHAQ: 0–3, pain VAS: 0–10, PtGA VAS: 0–10	Patient questionnaire	Patient: ~1.5 min	Provider: <30 sec Lab: N/A	Remission: 0–1.0 Low: 1.1–2.0 Moderate: 2.1–4.0 High: 4.1–10	Acceptable, test–retest reliability for composite has not been evaluated	Good	Acceptable for individual component, has not been evaluated for composite	Has been used in all rheumatic diseases Has demonstrated similar ability as compared to joint counts in differentiating active and inactive RA Includes patient-derived data only	No provider-derived data No acute-phase reactant included May be influenced by patient education level MDHAQ has a greater floor effect than original HAQ and less even spacing of questions than HAQII Requires raw scores to be converted to 0–10 prior to addition of components and division by 3
RAPID4PTJC	Disease activity measurement in rheumatic disease MDHAQ: 0–3, pain VAS: 0–10, PtGA VAS: 0–10, self-TJC: 0–48	Patient questionnaire, patient joint count	Patient: 5–10 min	Provider: <1 min Lab: N/A	Remission: 0–1.0 Low: 1.1–2.0 Moderate: 2.1–4.0 High: 4.1–10	Acceptable, test–retest reliability for composite has not been evaluated	Good	Acceptable for individual component, has not been evaluated for composite	Includes patient-derived data only	No provider-derived data No acute-phase reactant included May be influenced by patient education level MDHAQ has a greater floor effect than original HAQ and less even spacing of questions than HAQII Requires raw scores to be converted to 0–10 prior to addition of components and division by 4 Lacks standardization
RAPID4MDJC	Disease activity measurement in rheumatic disease MDHAQ: 0–3, pain VAS: 0–10, PtGA VAS: 0–10, TJC: 0–28, SJC: 0–26	Patient questionnaire, provider joint count	Patient: <1.5 min	Provider: ~2 min Lab: N/A	Remission: 0–1.0 Low: 1.1–2.0 Moderate: 2.1–4.0 High: 4.1–10	Acceptable, test–retest reliability for composite has not been evaluated	Good	Acceptable for individual component, has not been evaluated for composite	Includes patient- and provider-derived data	MDHAQ has a greater floor effect than original HAQ and less even spacing of questions than HAQII No acute-phase reactant included Requires raw scores to be converted to 0–10 prior to addition of components and division by 4 Lacks standardization
RAPID5	Disease activity measurement in rheumatic disease MDHAQ: 0–3, pain VAS: 0–10, PtGA VAS: 0–10, PrGA VAS 0–10, TJC: 0–48	Patient questionnaire, provider item, patient or provider joint count	Patient: 5–10 min if patient joint count and <5 min if provider joint count	Provider: <1 min if patient joint count and <2.5 min if provider joint count Lab: N/A	Remission: 0–1.0 Low: 1.1–2.0 Moderate: 2.1–4.0 High: 4.1–10	Acceptable, test–retest reliability for composite has not been evaluated	Good	Acceptable for individual component, has not been evaluated for composite	May include patient- and provider - derived data	No acute-phase reactant included MDHAQ has a greater floor effect than original HAQ and less even spacing of questions than HAQII Requires raw scores to be converted to 0–10 prior to addition of components and division by 5 Lacks standardization

(continued)

Summary Table *(Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden†	Score interpretation	Reliability evidence‡	Validity evidence‡	Ability to detect change‡	Strengths	Cautions
RADAI	RA disease activity measurement, may be useful in other forms of inflammatory arthritis Global activity over 6 months: 0–10, current activity: 0–10, pain: 0–10, health status: 0–6, self-TJC 0–48	Patient questionnaire	Patient: <5 min	Provider: <1 min Lab: N/A	Proposed remission/low: 0–2 Other categories of disease activity not established Higher values indicate poorer outcomes	Excellent	Good	Good	Includes patient-derived data only	No acute-phase reactant included No provider-derived data Not validated in non-RA disease states Validated only in whites
RADAI-5	RA disease activity measurement, may be useful in other forms of inflammatory arthritis Global activity over 6 months: 0–10, current activity: 0–10, pain: 0–10, health status: 0–10, stiffness: 0–10	Patient questionnaire	Patient: <1 min	Provider: <30 sec Lab: N/A	Remission: 0.0–1.4 Low: 1.6–3.0 Moderate: 3.2–5.4 High: 5.6–10.0	Good, test–retest reliability for composite has not been evaluated	Good	Poor	Includes patient-derived data only	No acute-phase reactant included No provider-derived data Not validated in non-RA disease states Validated only in whites
CASI	RA disease activity measurement HAQ: 0–3, pain VAS: 0–100, ESR: 0–100, RAI: 0–78	Patient questionnaire, provider assessment, lab	Patient: <3.5 min	Provider: ~3 min Lab: 1 hour waiting time for ESR	Remission: ≤24.65 Additional categories of disease activity not established	Excellent	Good	Good	Inclusion patient- and provider-derived data and a laboratory marker of inflammation	Most validation studies performed on measures no longer in general use Requires calculator/ computer for calculation
MOI-RA	RA disease activity measurement HAQ: 100 (0–3), PtGA VAS: 0–100, pain VAS 0–100, PrGA VAS: 0–100, ESR: 0–100	Provider item, patient item, lab	Patient: 5–10 min	Provider: 3–8 min Lab: 1 hour waiting time for ESR	Categories of disease activity not established Higher values indicate poorer outcomes	Good, test–retest reliability for composite has not been evaluated	Good, studies on predictive validity have not been done	Excellent	Includes all 7 ACR core set measures	Has not been validated outside inception cohort Lacks standardization
PDAS1	RA disease activity measurement MHAQ and self-TJC: 0–50, PtGA VAS: 0–10, ESR: 0–100	Patient questionnaire, lab	Patient: 5–10 min	Provider: <1 min Lab: 1 hour waiting time for ESR	Categories of disease activity not established Higher values indicate poorer outcomes	Acceptable, test–retest reliability for composite has not been evaluated	Acceptable	Acceptable for individual components	Self-administered Includes acute-phase reactan	Requires calculator/ computer for calculation Patient-derived joint counts Uses MHAQ, which has more pronounced floor effect than original HAQ, MDHAQ, or HAQII Not validated for other rheumatic disorders May be influenced by patient education level

(continued)

Summary Table *Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden†	Score interpretation	Reliability evidence‡	Validity evidence‡	Ability to detect change‡	Strengths	Cautions
PDAS2	RA disease activity measurement MHAQ and self-SJC: 0–28, PtGA VAS: 0–10, stiffness: 0–5	Patient questionnaire	Patient: 5–10 min	Provider: <1 min Lab: N/A	Higher values indicate poorer outcomes	Acceptable, test–retest reliability for composite has not been evaluated	Acceptable	Acceptable for individual components	Self-administered Measures morning stiffness	Requires calculator/ computer for calculation Patient-derived joint counts Uses MHAQ, which has more pronounced floor effect than original HAQ, MDHAQ, or HAQII No acute-phase reactant included Not validated for other rheumatic disorders May be influenced by patient education level
PtGA VAS	Activity measurement for any disease state 0–10 or 0–100 VAS	Patient item	Patient: ~10 sec	Provider: ~10 sec Lab: N/A	Categories of disease activity not established Higher values indicate poorer outcomes Proposed low: ≤2.0	Good	Acceptable	Acceptable	Self-administered All aspects of disease important to patient considered	Lacks provider data No acute-phase reactant included May be influenced by patient education level May include permanent damage May include non-RA disease
PrGA VAS	Activity measurement for any disease state 0–10 or 0–100 VAS	Provider item	Patient: N/A	Provider: ~10 sec Lab: N/A	Categories of disease activity not established Higher values indicate poorer outcomes Proposed low: ≤1.5 (0–10)	Excellent	Good	Acceptable	Provider administered All aspects of disease important to provider considered	Lacks patient data May include permanent damage May include non-RA disease

* RA = rheumatoid arthritis; DAS = Disease Activity Score; RAI = Ritchie Articular Index (53 joints in 26 graded units, graded for tenderness on pressure; maximum score 78); SJC = swollen joint count; ESR = erythrocyte sedimentation rate; PtGA = patient global assessment of disease activity; VAS = visual analog scale; DAS28 = Disease Activity Score with 28-joint counts; TJC = tender joint count; SDAI = Simplified Disease Activity Index; CRP = C-reactive protein; CDAI = Clinical Disease Activity Index; PrGA = provider global assessment of disease activity; N/A = not applicable; PAS = Patient Activity Score; HAQ = Health Assessment Questionnaire; PASII = Patient Activity Score-II; HAQII = Health Assessment Questionnaire-II; RAPID2 = Routine Assessment of Patient Index Data with 2 measures; RAPID3 = Routine Assessment of Patient Index Data with 3 measures; MDHAQ = Multidimensional Health Assessment Questionnaire; RAPID4PTJC = RAPID4-patient joint count; RAPID4MDJC = RAPID4-provider joint count; RAPID5 = Routine Assessment of Patient Index Data with 5 measures; RADAI = Rheumatoid Arthritis Disease Activity Index; RADAI-5 = Rheumatoid Arthritis Disease Activity Index-5; stiffness = early morning stiffness scale; CASI = Chronic Arthritis Systemic Index; MOI-RA = Mean Overall Index for Rheumatoid Arthritis; ACR = American College of Rheumatology; PDAS1 = Patient-Based Disease Activity Score with ESR; MHAQ = Modified Health Assessment Questionnaire; PDAS2 = Patient-Based Disease Activity Score without ESR.

† Times were derived from multiple sources (authors of this article, published comments, and formal studies) in varying populations and should be considered approximate estimates. For the most part, they were not determined in head-to-head studies, precluding direct comparison between measures (7,86,88,101,121).

‡ Psychometric properties were subjectively ranked using an ordered category scale (excellent, good, acceptable, poor, and unacceptable) based on currently published literature.

Summary Table for Disease Activity Measures in Rheumatoid Arthritis*

	Assessment	Require laboratory test	Physician joint count	Self-joint count	Patient global VAS	Provider global VAS	HAQ version	Morning stiffness	Pain	Defined remission criteria
DAS	Provider and patient	Yes	Yes	No	Yes	No	N/A	No	No	Yes
DAS28	Provider and patient	Yes	Yes	No	Yes	No	N/A	No	No	Yes
SDAI	Provider and patient	Yes	Yes	No	Yes	Yes	N/A	No	No	Yes
CDAI	Provider and patient	No	Yes	No	Yes	Yes	N/A	No	No	Yes
PAS	Patient	No	No	No	Yes	No	HAQ	No	Yes	Yes
PASII	Patient	No	No	No	Yes	No	HAQII	No	Yes	Yes
RAPID2	Provider and patient	No	No	No	Yes	Yes	N/A	No	No	No
RAPID3	Patient	No	No	No	Yes	No	MDHAQ	No	Yes	Yes
RAPID4PTJC	Patient	No	No	Yes	Yes	No	MDHAQ	No	Yes	Yes
RAPID4MDJC	Provider and patient	No	Yes	No	Yes	No	MDHAQ	No	Yes	Yes
RAPID5	Provider and patient	No	No	Yes	Yes	Yes	MDHAQ	No	Yes	Yes
RADAI	Patient	No	No	Yes	Yes	No	N/A	Yes	Yes	No
RADAI-5	Patient	No	No	No	Yes	No	N/A	Yes	Yes	Yes
CASI	Provider and patient	Yes	Yes	No	No	No	HAQ	No	Yes	Yes†
MOI-RA	Provider and patient	Yes	Yes	No	Yes	Yes	HAQ	No	Yes	No
PDAS1	Patient	Yes	No	Yes	Yes	No	MHAQ	No	No	No
PDAS2	Patient	No	No	Yes	Yes	No	MHAQ	Yes	No	No
PtGA	Patient	No	No	No	Yes	No	N/A	No	No	Yes†
PrGA	Provider	No	No	No	No	Yes	N/A	No	No	Yes†

* VAS = visual analog scale; HAQ = Health Assessment Questionnaire; DAS = Disease Activity Score; N/A = not applicable; DAS28 = Disease Activity Score with 28-joint counts; SDAI = Simplified Disease Activity Index; CDAI = Clinical Disease Activity Index; PAS = Patient Activity Score; PASII = Patient Activity Score-II; HAQII = Health Assessment Questionnaire-II; RAPID2 = Routine Assessment of Patient Index Data with 2 measures; RAPID3 = Routine Assessment of Patient Index Data with 3 measures; MDHAQ = Multidimensional Health Assessment Questionnaire; RAPID4PTJC = RAPID4-patient joint count; RAPID4MDJC = RAPID4-provider joint count; RAPID5 = Routine Assessment of Patient Index Data with 5 measures; RADAI = Rheumatoid Arthritis Disease Activity Index; RADAI-5 = Rheumatoid Arthritis Disease Activity Index-5; CASI = Chronic Arthritis Systemic Index; MOI-RA = Mean Overall Index for Rheumatoid Arthritis; PDAS1 = Patient-Based Disease Activity Score with erythrocyte sedimentation rate (ESR); MHAQ = Modified Health Assessment Questionnaire; PDAS2 = Patient-Based Disease Activity Score without ESR; PtGA = patient global assessment of disease activity; PrGA = provider global assessment of disease activity.

† While remission criteria for CASI, PtGA, and PrGA have been proposed, additional levels of disease activity are not defined.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S174–S188
DOI 10.1002/acr.20630

Measures of Adult Shoulder Function

Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and Its Short Version (*Quick*DASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society Standardized Shoulder Assessment Form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI)

FELIX ANGST,[1] HANS-KASPAR SCHWYZER,[2] ANDRÉ AESCHLIMANN,[3] BEAT R. SIMMEN,[2] AND JÖRG GOLDHAHN[2]

INTRODUCTION

There exists a large number of instruments that measure symptoms and function of the shoulder. More than 30 different tools can be found by entering “shoulder” and “assessment” into PubMed and conducting a review of the ≥3,000 retrieved references. Literature for every instrument was systematically reviewed by the key words “shoulder” and “instrument’s name.” We selected those that are cited in at least 20 references and for which psychometric testing has been reported. For each of these 9 tools, the 10–20 most informative studies about psychometric results were selected for citation to limit the references’ lists, but the entire body of literature was reviewed.

The Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH), together with its short form (*Quick*DASH), is the most widespread and best-tested and characterized instrument for shoulder assessment. However, it is region specific, i.e., specific to the arm, not just to the shoulder. The DASH stands out as an instrument positioned between the generic (as, for example, the Short Form 36) and the shoulder-specific measures, i.e., all other tools of the review: it forms the link between these 2 philosophies. It is a must for comprehensive assessment in conditions affecting different regions of the arm and for research studies. This review was focused only on shoulder studies of the DASH/*Quick*DASH.

The Shoulder Pain and Disability Index (SPADI), the Constant (Murley) Score (CS), and the American Shoulder and Elbow Surgeons (ASES) questionnaire for the shoulder are also well characterized and accepted in the scientific community. Their responsiveness is comparable. The SPADI is, together with the patient ASES, the shortest self-assessment and shows high validity. The ASES is a sophisticated measure for the patient and the examiner offering a relatively large number of items, often too long for clinicians. There are sparse data about the clinical (examiner-based) part of the ASES. The CS is the shortest self- and examiner-based tool. It combines the data of both into 1 total score. However, its intertester reliability is low and its validity is affected by the problem of different protocols on how to measure strength.

The Simple Shoulder Test (SST) is very short, very easy to understand and to score, and widely used in US. The binary item-response options (yes/no) affect the usability of the SST as metric score, validity, and comparability to other scores; the same is true for the Shoulder Disability Questionnaire (SDQ). The Oxford Shoulder Score was developed specifically for surgical conditions and is often used in the UK. It is very short, but there is a lack of psychometric testing data. The SDQ is very short but cannot be recommended due to absence of data on or weakness of psychometric properties.

Finally, the Western Ontario Shoulder Instability index (WOSI) was selected because, in the last few years, it has become the most often used and best psychometrically tested assessment of shoulder instability, although there is still a lack of testing data.

For a set of clinical assessment tools, we recommend the *Quick*DASH, the SPADI (or the patient ASES), and the CS, and the WOSI if instability is part of the condition. For a research set, the DASH, the SPADI, and, possibly, the clinical part of the ASES or the CS can be recommended in order to (also) obtain more information about examiner-based data.

[1]Felix Angst, MD, MPH: RehaClinic Zurzach, Bad Zurzach, and Schulthess Klinik, Zurich, Switzerland; [2]Hans-Kaspar Schwyzer, MD, Beat R. Simmen, MD, Jörg Goldhahn, MD: Schulthess Klinik, Zurich, Switzerland; [3]André Aeschlimann, MD: RehaClinic Zurzach, Bad Zurzach, Switzerland.

Address correspondence to Felix Angst, MD, MPH, RehaClinic Zurzach, Quellenstrasse, 5330 Bad Zurzach, Aargau, Switzerland. E-mail: fangst@vtxmail.ch.

Submitted for publication January 31, 2011; accepted in revised form May 10, 2011.

DISABILITIES OF THE ARM, SHOULDER, AND HAND QUESTIONNAIRE (DASH) AND ITS SHORT VERSION (*QUICK*DASH)

Description

Purpose. Self-assessment of symptoms and function of the entire upper extremity (1).

Settings. All domains, any or multiple disorders of the upper extremity.

Versions. Original version (30 items) and derivations of it as short versions (11 or 9 items); preliminary publication in 1996 (2), first publication of the manual in 1999, second edition in 2002, and third edition in 2011 (1); *Quick*DASH in 2005 (3); and *Quick*DASH-9 in 2009 (4).

Content and number of items. 30 items (total score): 6 items about symptoms (3 about pain, 1 for tingling/numbness, 1 for weakness, 1 for stiffness) and 24 about function (21 about physical function, 3 about social/role function). Determination of the subscores symptoms and function is possible, but this is not originally described (1,5–9). Two optional additional modules for work (4 items) and sports/performing arts (4 items) are more rarely used in patient settings, but rather for manual workers and athletes. The "classic" *Quick*DASH has 11 items (3 for symptoms, 8 for function) and will be referred to throughout as the "*Quick*DASH" (3,10,11). Other short versions exist, e.g., the *Quick*DASH-9 (1 item for pain, 8 for function), but are rarely used and not supported by the authors of the original (1,4).

Response options/scale. All items are scored on a scale of 5 (Likert) levels: 1 = no difficulty/symptoms, 2 = mild difficulty/symptoms, 3 = moderate difficulty/symptoms, 4 = severe difficulty/symptoms, and 5 = extreme difficulty (unable to do)/symptoms.

Recall period for items. 1 week.

Endorsements. American Association of Orthopedic Surgeons and Institute of Work and Health (IWH) (1).

Examples of use. Relevant settings (aims and analysis [references]) for the DASH are as follows:

Various regions of upper extremity (development of the DASH [2])
Various regions of upper extremity (DASH manual: third edition [1])
Various regions of upper extremity (population normative data [1,12])
Shoulder instruments (important comparative reviews [7,13])
Various regions of upper extremity (reliability, validity, responsiveness [14])
Various operations of upper extremity (responsiveness [15])
Various regions of upper extremity (validity, factor, Rasch [9])
Upper extremity, neck pain (validity, responsiveness [16])
Upper extremity, lower extremity (validity [17])
Rheumatoid arthritis (reliability, validity [18])
Multiple sclerosis (reliability, validity, Rasch [8])
Shoulder arthroplasty (responsiveness [19])
Adhesive capsulitis (validity, responsiveness [20])
Shoulder impingement, tendinitis (validity, responsiveness, minimum clinically important difference [MCID] [21])
Proximal humerus fracture (reliability, validity [22])
Elbow, arthroplasty (validity [23,24])
Distal radius facture (reliability, validity, responsiveness [25])
Hand osteoarthritis, fractures (responsiveness [26])
Hand, various (validity, German DASH [5])
Rhizarthrosis (validity [27])

Relevant settings (aims and analysis [references]) for the *Quick*DASH are as follows:

Various regions of upper extremity (development of the *Quick*DASH [3])
Various surgery of upper extremity (psychometric testing of the *Quick*DASH [10,11])
Shoulder pain (reliability, MCID [28])
Various regions of upper extremity (development of the *Quick*DASH-9 [4])

Practical Application

How to obtain. Property and copyright at the IWH (online at http://www.dash.iwh.on.ca/). There, further links lead to the forms for free for the DASH (http://www.dash.iwh.on.ca/assets/images/pdfs/dash_questionnaire_2010.pdf) and *Quick*DASH (http://www.dash.iwh.on.ca/assets/images/pdfs/quickdash_questionnaire_2010.pdf). Language versions are online at http://www.dash.iwh.on.ca/translate.htm. Free of charge for noncommercial use; license for commercial use available at the IWH. Manual (3rd edition) online and paper copy; costs not yet determined.

Method of administration. Self-assessment.

Scoring. The arithmetic mean of at least 27 of the 30 items (missing rule) is transformed by (mean − 1) × 25 into the scale from 0 = no symptoms/full function to 100 = maximal symptoms/no function for the DASH total score (1,11). Five of 6 items are necessary for determination of the symptoms score and 22 of 24 items for the function score (11). Similarly, 10 of 11 items are necessary for the *Quick*DASH total score, 3 of 3 for symptoms, and 7 of 8 for function (3,10,11). Computer scoring is not necessary but easier, e.g., on Microsoft Excel or any calculation or statistics program. Scoring program is online at http://www.dash.iwh.on.ca/score.htm.

Score interpretation. Originally, 0 = best and 100 = worst. The reverse scale from 0 = worst to 100 = best by (100 = original score) is also often used for comparison with other scores, e.g., the Short Form 36 (SF-36). Several studies showed varying distinct cutoff points to reflect severity (1). Cutoff scores: <15 = "no problem," 16–40 = "problem, but working," and >40 = "unable to work" (1). Normative values of 1,706 persons in the US general population, stratified by sex, age, and comorbidity, are available (US population mean ± SD 10.1 ± 14.7) (1,6,12).

Respondent burden. Time to complete is 4 minutes for the DASH and 2 minutes for the *Quick*DASH (1,3,6,7). All items are easy to comprehend and are not emotionally sensitive (with the exception of item 21; see below).

Administrative burden. Item rating can be typed or scanned into an electronic database. Score computation is easy (see above). The head of the questionnaire contains instructions on how to complete it. Time to administer (including control of missing data): DASH, 10 minutes; *Quick*DASH, 8 minutes (1). Time to scan and determine the scores: 2 minutes. Little special training is necessary for these activities.

Translations/adaptations. Available for free for 35 languages and dialects. Versions in 11 other languages are in progress (as of January 30, 2011).

Psychometric Information

Method of development. Eight hundred twenty-one possible questions obtained by literature review were reduced to 67 (+3 new) due to content overlap or off target by a consensus group. Patient data were analyzed by different item to total correlation techniques, comparison to clinimetric ranking, and clinical judgment, resulting in the final 30-item version (1,2). The newest manual contains extensive psychometric information (1). Psychometric analysis by item-response theory (using Rasch analysis) was performed later for the DASH (8,9). All relevant modern strategies were used in the development of the *Quick*DASH comparing 3 strategies: the concept-retention method, the equidiscriminative item-total correlation, and the item-response theory (Rasch modeling). The concept-retention method was most similar to the DASH and was chosen to build the *Quick*DASH (3).

Acceptability. All item content is easy to read and understand. Missing data are rare. Item 21 that asks about sexual activity is often left out by patients. For that reason, item 21 has been skipped in the *Quick*DASH (3,6). Low floor and ceiling effects are reported (1,6,8,11,14,18).

Reliability. Internal consistency/cross-sectional reliability: Cronbach's $\alpha = 0.92–0.98$ for the DASH (1,4,8,9,15) and 0.92–0.95 for the *Quick*DASH (3,10).

Test–retest reliability: intraclass correlation coefficient 0.93–0.98 for the DASH (1,14,18,21,22) and 0.90–0.94 for the *Quick*DASH (3,10,28).

Validity. *Content validity.* Normally distributed scores and low floor and ceiling effects (6,14,18).

Criterion validity. There is no gold standard for symptoms or function measurement of the shoulder. The obvious content validity of the used items and the numerous studies of the DASH give it a certain intrinsic validity. However, criterion validity of the DASH came into question when Rasch analysis was applied (8,9). The corresponding results for the *Quick*DASH were better but also criticized (3,9).

Construct validity. Pearson's or Spearman's correlations of the DASH total score to other instruments are as follows:
SPADI: 0.79–0.93 and 0.55 (ref. 6,14,20)
HAQ: 0.88 and 0.54 (ref. 18,20)
CS: 0.82 (ref. 6)
ASES: 0.79 (ref. 6)
EQ-5D: 0.75 (ref. 22)
SF-12 PCS: 0.75 and 0.57–0.63 (ref. 16,22)
SF-36 PCS: 0.70 (ref. 6,18)
Global disability rating: 0.67–0.71 (ref. 21)
DAS28: 0.42 (ref. 18)
SF-36 MCS: 0.27 and 0.06 (ref. 6,18)
SF-12 MCS: 0.10–0.33 (ref. 16)

The correlations reflected a well-fitting dose-response curve for the construct of shoulder specificity of the compared instruments (19). Extraordinary low correlations were reported in 1 study (20).

Pearson's correlations of the *Quick*DASH total score to other instruments are as follows:
SPADI: 0.84 (ref. 11)
SF-36 PCS: 0.68 (ref. 11)
Global rating of change: 0.45 (ref. 28)

Ability to detect change. Minimally detectable change (MDC95%): 7.9–14.8 points for the DASH (7,14,21) and 13.3 for the *Quick*DASH (28).

MCID: 10.2 points (21). Comparison and critique of different methods to determine MCID on the DASH was done (29). *Quick*DASH: 8.0 points (28). Between-group differences are reported (1,7).

Effect sizes (ES) and standardized response means (SRMs) of the DASH total score in shoulder conditions are as follows:
Total shoulder arthroplasty: ES 1.19, SRM 1.22 (ref. 19)
Neck and/or shoulder at general practitioner: ES 0.88–0.90, SRM 0.88–0.93 (ref. 16)
Arthroscopic acromioplasty: ES 0.9, SRM 0.5 (ref. 15)
Neck symptoms at general practitioner: ES 0.88, SRM 0.88 (ref. 16)
Shoulder impingement, tendinitis: physiotherapy: ES 0.81, SRM 0.72 (ref. 21)
Rotator cuff surgery, total shoulder arthroplasty: ES 0.64, SRM 0.81 (ref. 14)
Adhesive capsulitis: steroids: ES 0.34, SRM 0.43 (ref. 20)

ES and SRMs of the *Quick*DASH total score in shoulder conditions are as follows:
Total shoulder arthroplasty: ES 1.26 (ref. 11)
Shoulder or hand: conservative treatment: SRM 0.79 (ref. 3)
Various upper extremity surgery: ES 0.50, SRM 0.63 (ref. 10)

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The DASH is the best-tested and most often used self-assessment instrument for the shoulder and any other disorders of the upper extremity. It is particularly useful in polyarticular conditions or if measurement of symptoms and function of the entire upper extremity is wanted. Since shoulder function determines the position of the elbow and the hand, the DASH is also useful in all elbow and hand conditions. Some of the DASH items also ask about fine-motor hand functions. Empiric data can be compared to US population norms. The *Quick*DASH total score yields very similar values to those of the DASH and the total scores correlate highly to each other (3,11).

Caveats and cautions. The DASH is region specific, not joint specific. Specificity and responsiveness of the DASH are, therefore, lower than those of unique shoulder-specific tools but higher than those of generic quality of life tools (19). Compared to other instruments, the strict 90% missing rule produces a relatively high percentage of missing data. There is evidence that the DASH score is also influenced by disability of the lower extremity (17). Rasch analysis revealed problems with the unidimensionality of the DASH total score and with differentiation between "mild/moderate/severe difficulty," which affects (criterion) validity (8,9). Obvious misfits were items 21 (sexual activity) and 26 (tingling) (3,8,9). Item 26 is retained in the *Quick*DASH. However, this needs closer investigation as a classically developed tool is fitted into a modern measurement framework. The *Quick*DASH has a similar total score to the DASH but it underestimates symptoms (reports lower severity) and overestimates function (reports less disability) when compared to the DASH (11). In the case where an MDC95% is reported to be higher than the MCID, the MDC95% should be taken as the MCID.

Clinical usability. The DASH is the best tool for comprehensive assessment of upper extremity conditions, e.g., if shoulder problems cannot be differentiated from hand problems (rheumatoid arthritis, polytrauma, multiple sclerosis). It is easy to apply, analyze, and interpret. Comparison of empirical and normative data allows valid description of the patient's upper extremity status. The *Quick*DASH provides the necessary short assessment for clinical visits.

Research usability. The DASH is good for research purposes in various upper extremity conditions. It is well tested and there is a large body of data for comparison of different settings and different upper extremity instruments, especially for analysis of construct validity compared to other instruments. The concerns about validity obtained by Rasch analysis cannot be disregarded, but development of new methods to assess validity, e.g., item-response theory, is ongoing. Specificity and responsiveness in localized conditions (affecting only 1 joint) are moderate. The use of the subscales symptoms and function are recommended for the DASH but not for the *Quick*DASH (11). The constructs of the 2 instruments are not exactly the same.

SHOULDER PAIN AND DISABILITY INDEX (SPADI)

Description

Purpose. Self-assessment of symptoms and function of the shoulder.

Settings. All domains, any disorders of the shoulder joint.

Versions. Original version published in 1991 (30). No revisions.

Content and number of items. 13 items (total score): 5 items for pain and 8 for function (subscores).

Response options/scale. All SPADI items are originally scored on a visual analog scale (VAS) from no pain/no difficulty to worst pain imaginable/so difficult required help. The VAS line was divided into 12 equal intervals to obtain a 12-point numerical rating scale (NRS) ranging from 0 (best) to 11 (worst) (30). Later versions used the 12-point or an 11-point NRS (0–10) without a VAS line (31).

Recall period for items. 1 week.

Endorsements. None.

Examples of use. Relevant settings (aims and analysis [references]) for the SPADI are as follows:

Shoulder pain (development of the SPADI [30])
Shoulder instruments (important comparative reviews [7,13])
Various upper extremity diagnoses (reliability, minimal detectable difference [MDD], minimum clinically important difference [MCID] [21])
Various shoulder diagnoses (validity [32])
Adhesive capsulitis (factor analysis [33])
Adhesive capsulitis (reliability, validity, responsiveness [20,34])
Rotator cuff (reliability, validity [35])
Rotator cuff, local infiltration (MCID [36])
After shoulder arthroplasty (validity, MDC [6,31])
Total shoulder arthroplasty (responsiveness [19])
Various shoulder surgery (reliability, responsiveness [37])
Orthopedic practice (validity, factor, MDC, MCID [38])
Orthopedic practice (Rasch, partial credit model [39])
Primary care (validity, responsiveness [40])
Outpatient physiotherapy (validity, responsiveness [41])
Community volunteers (factor analysis [42])

Practical Application

How to obtain. Printed in various references (30,31,40–42). Free online at http://www.workcover.com/site/treat_home/outcome_measures_and_risk_screening_tools/links_to_outcome_measures_and_screening_tools.aspx?.

Method of administration. Self-assessment.

Time to complete. 2–3 minutes (7,37).

Scoring. Originally, the sum of marked items/maximal possible score × 100 with at least 11 of 13 completed items necessary for the total score (30). Later and with permission of the developer K. E. Roach, the "2/3 missing rule," as used for many instruments, was applied: at least 3 of 5 pain and 6 of 8 function items for the subscales are necessary (6,31). The SPADI total score is the unweighted mean of the pain and function subscores (30). In fact, the (sub)scores can be determined by the arithmetic mean of the completed items by mean/11 × 100 using the 12-point NRS (or mean × 10 using the 11-point NRS). Computer scoring is not necessary but easier.

Score interpretation. Originally, 0 = best and 100 = worst. A reverse scale from 0 = worst to 100 = best (100 = original score) is also often used to compare with other scores, e.g., the Short Form 36 (SF-36). There are no distinct cutoff points to reflect severity. Empirical normative values are not determined.

Respondent burden. All items are easy to comprehend and are not emotionally sensitive.

Administrative burden. Score computation is easy. The head of the questionnaire contains a short explanation on how to complete it. Time to administer: 5 minutes (30). Time to scan and determine the scores: 2 minutes.

Translations/adaptations. Published in 3 languages: Norwegian (34), German (31), and Slovene. Versions in Chinese, Hindi, Brazilian Portuguese, Japanese, Turkish, and French Canadian exist but have not been published under peer review (Roach KE: unpublished observations).

Psychometric Information

Method of development. 20 items were selected by a group of 3 rheumatologists and 1 physiotherapist and established by assessing their face validity for pain and function, their test–retest reliability, and their correlation to shoulder range of motion (30). Item-response theory was applied to the function subscale only (39).

Acceptability. Easy to read and understand. Missing data are very rare. Low floor and ceiling effects reported (6,31,32,41).

Reliability. Internal reliability/consistency: Cronbach's $\alpha = 0.86-0.96$ (30,31,33,38,40,42).

Test–retest reliability: intraclass correlation coefficient 0.84–0.95 (7,21,31,34,37). It was exceptionally low with 0.66 in the development study (30).

Validity. *Content validity.* The scores were normally distributed in 1 study (6) but not in 2 studies (31,41). Low floor and ceiling effects were seen, especially for the function subscore (6,31,32,41).

Criterion validity. In the absence of a gold standard, the obvious content validity of the used items and the numerous studies examining the SPADI give it a certain intrinsic validity. Rasch and factor analysis revealed moderate overall criterion validity: items 8 (removing something from the back pocket), 7 (carrying ≥10 lbs), and 4 (closing front buttons) showed some misfit (only the function subscore was examined) (39). Very low and very high function were not precisely measured (39). The 2 subscores pain and function could not be supported by factor analysis (33,38,42).

Construct validity. Pearson's or Spearman's correlations of the SPADI total score to other instruments are as follows:

DASH: 0.93, 0.55, and 0.88 (ref. 6,20,31)
ASES: 0.81, 0.92, and 0.77 (ref. 6,31,37)
OSS: 0.57 and 0.85 (ref. 35,43)
CS: 0.82 (ref. 6)
SST: 0.74 and 0.80 (ref. 32,38)
SF-36 PCS: 0.63 and 0.67 (ref. 6,32)
Global disability rating: 0.64–0.69 (ref. 21)
HAQ: 0.55 and 0.61 (ref. 20,40)
Sickness Impact Profile: 0.57 (ref. 41)
Active ROM: 0.54–0.80 and 0.38 (ref. 30,34)
SF-36 MCS: 0.08 (ref. 6)
Extraordinary low correlations were reported in 1 study (20).

Ability to detect change. Minimally detectable change (MDC95%) for the total score: 17.0, 13.2, 17.2, and 21.5 points, respectively, as calculated in 4 studies (21,31,34,38).

MCID: 13.2, 15.4, and 23.1 points, respectively (21,36,38).

Effect sizes (ES) and standardized response means (SRMs) of the SPADI total score are as follows:

Total shoulder arthroplasty: ES 2.10, SRM 1.72 (ref. 19)
Adhesive capsulitis: steroids: ES 1.94, SRM 1.81 (ref. 34)
Adhesive capsulitis: steroids: ES 1.20–1.64, SRM 1.27–1.68 (ref. 20)
Shoulder pain, physiotherapy: ES 1.26, SRM 1.38 (ref. 41)
Rotator cuff surgery + total shoulder arthroplasty: SRM 1.23 (ref. 37)
Various upper extremity, occupational, physiotherapy: ES 0.80, SRM 0.67 (ref. 21)
General practice, conservative therapy: ES 0.34 (ref. 40)

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SPADI is the most responsive shoulder instrument and has been tested in numerous settings. It is short; it is easy to understand, complete, and analyze; and no costs are involved in obtaining it.

Caveats and cautions. Criterion and construct validity showed some weaknesses in factor and Rasch analysis. The original 12-item NRS (where 0 = best and 11 = worst) is uncommon. Only 1 item assesses overhead work or heavy use of the shoulder, which may produce ceiling effects. In the case where an MDC95% is reported to be higher than the MCID, the MDC95% should be taken as the MCID.

Clinical usability. Very good for short and responsive assessment in all shoulder conditions. Easy to interpret.

Research usability. Most responsive shoulder tool (19,37). Recommended for every set of shoulder assessments. Subscores with limited criterion validity.

AMERICAN SHOULDER AND ELBOW SURGEONS (ASES) SOCIETY STANDARDIZED SHOULDER ASSESSMENT FORM

Description

Purpose. Developed to "represent a state-of-the-art questionnaire with three key features: 1) ease of use 2) method of assessing activities of daily living (ADL) and 3) inclusion of a patient self-evaluation section," approved by the ASES Research Committee in 1994 (44) to be applicable to all shoulder patients regardless of diagnosis. In 1998, the original ASES was modified to the mASES by deleting 2 and adding 5 function items to make a "whole-extremity questionnaire rather than a shoulder questionnaire alone" (37). This chapter deals with the original ASES only, not with the mASES.

Content and number of items. Patient self-assessment section (patient ASES [pASES]) and a section to be completed by the examiner (clinical ASES [cASES] or, more precisely, ASES-examiner). The pASES form is divided into 3 sections: pain (6 items), instability (2 items), and ADL (10 items for both sides each). The cASES has 4 parts (each for left and right): range of motion (5 items, each passive and active), signs (11 items), strength (5 items), and instability (8 items + 1 open question).

Response options/scale. Binary (yes/no) answers for pain and instability, visual analog scales (VAS) for pain and instability (where 0 = best and 10 = worst), and 4-point ordinal Likert scale for function (where 0 = unable to do, 1 = very difficult, 2 = somewhat difficult, and 3 = not difficult).

Recall period for items. 1 week.

Endorsements. ASES (44).

Examples of use. Relevant settings (aims and analysis [references]) for the ASES are as follows:

No empirical field testing (development of the ASES [44])
Outpatients without shoulder problems (normative data [45])
Shoulder instruments (important comparative reviews [7,13])

Various shoulder dysfunctions (reliability, validity, responsiveness [46])
Subacromial impingement (validity [47])
Calcific tendinitis (responsiveness [48])
Rotator cuff, tendinitis (minimum clinically important difference [MCID] [49])
Rotator cuff, arthritis (Italian ASES, reliability, validity [50])
Rotator cuff, instability, arthritis (reliability, validity, responsiveness [51])
Rheumatoid arthritis, osteoarthritis (German ASES, reliability, validity [52])
Orthopedic practice (reliability [53])
Osteoarthritis, hemi- or total arthroplasty (responsiveness [54])
Total shoulder arthroplasty (validity, responsiveness [6,19])

Practical Application

How to obtain. Original publication (44). Free online at http://www.shoulderandkneesurgery.com/pdf/ases_assessment_form.pdf.

Method of administration. Self-assessment.

Time to complete. 3 minutes (pASES).

Scoring. The pASES total score = ((10 − VAS pain) × 5) + (5/3 × sum of ADL items) (44). The instability items and the remaining 5 pain items do not contribute to the pASES total score. Determination of the cASES was not described originally; 1 solution, using 2 of 3 of the completed items to determine the scores, is given in 1 study (6).

Score interpretation. 0 = worst and 100 = best. An original missing rule and distinct cutoffs to reflect severity have not been published. Normative data are provided in graph form, stratified by 10-year age groups but not by sex (45).

Respondent burden. Time to complete is 3 minutes for the pASES (44). All items are easy to understand and are not suggestive or emotionally sensitive.

Administrative burden. The patient section can be administered without the clinical section. This is short to perform and is done in most of the applications. Score computation is easy and can be implemented in any database. Time (pASES): 8 minutes (estimated). Patient examination for the cASES is time consuming.

Translations/adaptations. German (52), Italian (50), and Portuguese.

Psychometric Information

Method of development. Developed by the research committee of the ASES that reviewed existing instruments at that time through open discussion and without a specific methodologic approach.

Acceptability. All item content is easy to read and understand. Missing data are very rare. Single items may show high floor and ceiling effects (52).

Reliability. Internal reliability/consistency: Cronbach's α = 0.61–0.96 (46,50–53).

Test–retest reliability: intraclass correlation coefficient 0.84–0.96 (45,46,50–52).

Validity. *Content validity.* Content validity was questioned in 1 study (13). Minimal floor and ceiling effects of the total score are described in 2 studies (50,51), but higher ones are also described in 2 additional studies (6,52). Normal distribution of the scores is reported (6).

Criterion validity. In the absence of a gold standard, the obvious content validity of the used items and the numerous studies of the pASES give it a certain intrinsic validity. The ASES has not been examined by item-response theory, factor, or Rasch analysis.

Construct validity. Pearson's or Spearman's correlations of the pASES total score to other instruments are as follows:

SPADI: 0.92 and 0.81 (ref. 6,52)
Western Ontario Rotator Cuff index: 0.81 (ref. 47)
DASH: 0.79–0.92 (ref. 6,50,52)
CS: 0.71 (ref. 6)
Rotator Cuff QOL: 0.70 (ref. 47)
SF-36 bodily pain: 0.60 and 0.65 (ref. 50,52)
SF-36 PCS: 0.48 and 0.64 (ref. 6,50,52)
SF-36 physical functioning: 0.47 and 0.57 (ref. 50,52)
cASES: 0.48 (ref. 6)
SF-36 MCS: 0.24 and −0.20 (ref. 6,50)

Ability to detect change. Minimally detectable change (MDC95%): 11.2 (46).

Minimum clinically important difference (MCID): 6.4 (46) and 12.0–16.9 (49).

Effect sizes (ES) and standardized response means (SRMs) of the pASES total score are as follows:

Osteoarthritis: total or hemi shoulder arthroplasty: ES 3.53 (ref. 54)
Rheumatoid, osteoarthritis: total shoulder arthroplasty: ES 2.13, SRM 1.81 (ref. 19)
Calcific tendinitis: subacromial steroid: ES 1.65–1.84 (ref. 48)
Various, mainly impingement: physiotherapy: ES 1.39, SRM 1.54 (ref. 46)
Rotator cuff disease: SRM 1.42 (ref. 47)
Rotator cuff, instability, arthritis: surgery: ES 0.93–1.16 (ref. 51)

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Recommended by the ASES and, by that, widespread use, especially in American centers. The ASES showed good reliability, high construct validity, and high responsiveness.

Caveats and cautions. Mix of scales (binary, Likert, VAS). Limited content, especially criterion validity. In the case where an MDC95% is reported to be higher than the MCID, the MDC95% should be taken as the MCID.

Clinical usability. Helpful combination of self- and clinical assessment.

Research usability. Good applicability for research and good responsiveness. Slightly longer than and less frequently used as the Shoulder Pain and Disability Index. Some methodologic weaknesses.

CONSTANT (MURLEY) SCORE (CS)

Description

Purpose. "The method records individual parameters and provides an overall clinical functional assessment . . . applicable irrespective of details of the diagnostic or radiological abnormalities . . . , sufficiently sensitive to reveal even small changes in function" (55). Introduced in 1987 (55). Revision in 2008 (56).

Content and number of items. The score consists of 4 domains: pain (1 item), activities of daily living (ADL; 3 items for activity level, i.e., work, sports, sleep, 1 item for hand positioning, i.e., rotation), mobility (4 items: forward and lateral abduction/elevation, external and internal rotation), and power/strength (1 item). Pain and ADL 1–3 are interviewed from the patient (i.e., self-assessed); all other items are examiner assessed.

Response options/scale. Pain item: originally 4 Likert levels, visual analog scale in the revised version (55,56), where 0 = maximal pain and 15 = no pain. ADL: Likert scales, where 0 = worst and 5 = best for each item. Mobility: active, pain-free range of elevation: +2 points per 30°, where 0 = worst and 10 = best for each item; position of hand: 0 = worst to 10 = best (55–57). Strength is measured at 90° lateral abduction by use of either an Isobex device or a defined spring balance technique: 1 point per 0.5 kg (~1 lbs), maximum 25 points (56).

Recall period for items. 1 week.

Endorsements. European Society for Surgery of the Shoulder and the Elbow (SECEC-ESSSE) and recommend by the German Society of Shoulder and Elbow Surgeons.

Examples of use. Relevant settings (aims and analysis [references]) for the CS are as follows:

No empirical field testing (development of the CS [55])
Referring to previous studies (revision of the CS [56])
Systematic literature review (psychometric properties of the CS [57])
No shoulder pain/disability (normative CS values [58])
Various shoulder dysfunctions (intra- and intertester reliability [59])
Various, mainly rotator cuff (validity, responsiveness [60])
Impingement (validity, responsiveness [61–63])
Degenerative, inflammatory (validity [64])
Rotator cuff repair (validity [65,66])
Shoulder instability (validity, responsiveness [67])
Osteoarthritis (responsiveness [68])
Rheumatoid, osteoarthritis (validity, responsiveness [6,19])

Practical Application

How to obtain. Original publication (55) and online at http://www.secec.org/data/upload/files/Constant%20Score.pdf.

Method of administration. Clinical examination plus patient interview (self-assessment). Retrospective data extraction from the case history is not reliable, especially not for the patient's self-assessment items.

Time to complete. 5–7 minutes (61).

Scoring. The sum of the subscores results in the CS total score: pain (0–15) + ADL (4 × (0–5) = 0–20) + mobility (4 × (0–10) = 0–40) + strength (0–25).

Score interpretation. 0 = worst and 100 = best function. Comparison with the contralateral side is possible. Different norm data are available, and in the past, expressed as a percentage of age-adjusted norm data, the relative CS was recommended, but is problematic because of different norm cohorts (58).

Respondent burden. Minimal (see below). All items are easy to understand and not emotionally sensitive.

Administrative burden. Moderate because the CS can be implemented in a normal clinical investigation (57). The measurement of strength demands some extra effort. Score calculation is easy and can be implemented in any calculation software.

Translations/adaptations. The CS is used in almost every language without official translations because surgeons perceived the score as a clinical measure (57). In French, a validated translation/adaptation has been published.

Psychometric Information

Method of development. The score was originally developed as part of a master's thesis and later published (55). The methodology of development was not reported or specified. The score was revisited by the SECEC-ESSSE members (56).

Acceptability. High acceptance by patients because the items have a high relevance. Acceptance among surgeons is very high due to the clinical relevance.

Reliability. Internal reliability/consistency: Cronbach's α = 0.37 and 0.60, respectively (60,66).

Test–retest reliability: intraclass correlation coefficient 0.80–0.96 (57). Repeated strength measurements revealed high intratester but low intertester reliability (59).

Validity. *Content validity.* No floor and ceiling effects for the CS total score were shown, but the subscores, especially strength (when unable to reach 90° abduction), reached substantial floor levels (i.e., no strength) (6,64). The CS total score was normally distributed (6).

Criterion validity. There is no gold standard for self- and examiner-assessed shoulder function. There is an ongoing debate about the appropriate measure for abduction strength. Whereas originally an unsecured spring balance was utilized (55), the last modification of the score advocates Isobex measurement (56). However, both are strongly correlated to each other. Large variations in handling the testing protocol have been reported leading to a large interobserver variance (59). There are no data about factor, Rasch analysis, or item-response theory.

Construct validity. Pearson's or Spearman's correlations of the CS to other instruments are as follows:

ASES: 0.72–0.87 (ref. 6,65,66)
OSS: 0.65–0.87 (ref. 61,64)
DASH: 0.82, 0.76, and 0.50 (ref. 6,64)
SPADI: 0.53 and 0.82 (ref. 6,64)
WOSI: 0.58 (ref. 67)
SST: 0.49 (ref. 65)
SF-36 PCS: 0.45 (ref. 6)
Rating of change (shoulder): 0.32–0.70 (ref. 63)
SF-36 MCS: 0.02 (ref. 6)

Considerably low correlations were found in 1 study (64).

Ability to detect change. Minimally detectable change (MDC95%) and minimum clinically important difference (MCID): no data published.

Effect sizes (ES) and standardized response means (SRMs) of the CS total score are as follows:

Osteoarthritis: hemi or total shoulder arthroplasty: ES 3.02 (ref. 54)

Rheumatoid, osteoarthritis: total shoulder arthroplasty: ES 2.23, SRM 1.99 (ref. 19)

Impingement: arthroscopic decompression: ES 0.65–1.92, SRM 0.62–2.09 (ref. 63)

Impingement: open decompression: ES 1.60, SRM 1.39 (ref. 61)

Impingement: acupuncture, transcutaneous electrical nerve stimulation: ES 1.29 and 0.73 (ref. 62)

Shoulder instability: physiotherapy ± surgery: SRM 0.59 (ref. 67)

Various, mainly rotator cuff: surgery: ES 0.58, SRM 0.57 (ref. 60)

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CS covers the clinically most relevant domains and shows high responsiveness. It is highly accepted throughout the clinical community in the fields of arthroplasty, rotator cuff disease, shoulder trauma, and fractures.

Caveats and cautions. There are sparse, and in some parts, no data about reliability and validity (except construct validity). Intertester reliability was shown to be low. Different versions and measurement methodologies lead to problems when comparing data. How to measure strength has not been standardized yet. The relative CS (percentage of norm data) is invalid due to different norm data. Only 1 pain item and only 3 ADL items may be not sufficient to adequately assess self-rated pain and function. Due to lack of testing data (MDC95%, MCID), caution is necessary for measurement at an individual patient level.

Clinical usability. The CS is in widespread clinical use. The CS often serves as the mandatory part of a measurement protocol, especially in Europe. It is not suitable for patients with instability conditions. Due to lack of testing data or insufficient measurement properties, caution is necessary for measurement at an individual patient level.

Research usability. Limited due to the caveats, especially insufficiently testing of validity.

SIMPLE SHOULDER TEST (SST)

Description

Purpose. To assess functional disability of the shoulder (68).

Content and number of items. Total score of 12 items: 2 about function related to pain, 7 about function/strength, and 3 about range of motion (32). No subscales.

Response options/scale. Dichotomous responses: 1 = yes (function possible) and 0 = no.

Recall period for items. Actual/at the moment of assessment.

Endorsements. None.

Examples of use. Relevant settings (aims and analysis [references]) for the SST are as follows:

Normal and affected shoulders (development of the SST [68])

Shoulder instruments (important comparative review [7])

Various shoulder problems (validity, responsiveness [32,37])

Shoulder injuries (reliability, validity, responsiveness [69])

Shoulder joint destruction (responsiveness, minimum clinically important difference [MCID] [70])

Rotator cuff, conservative (MCID [49])

Rotator cuff repair (validity, responsiveness [71,72])

Orthopedic practice (validity, factor, minimal detectable difference [38])

Orthopedic practice (Rasch, partial credit model [39])

Practical Application

How to obtain. Original publication (68). Free online at http://www.orthop.washington.edu/PatientCare/OurServices/ShoulderElbow/Articles/SimpleShoulderTest.aspx.

Method of administration. Self-assessment.

Time to complete. 2–3 minutes.

Scoring. Original score: 0 = worst and 12 = best. Transformed by: number of "yes" items/number of completed items × 100 = % "yes" responses.

Score interpretation. 0 = worst and 100 = best function. A missing rule, distinct cutoffs for severity, and normative data have not been published.

Respondent burden. Very short; easy to understand and not emotionally sensitive.

Administrative burden. Free online. Score computation is very easy and possible by hand. Time to administer and determine: estimated 5 minutes.

Translations/adaptations. No data published.

Psychometric Information

Method of development. "Questions derived from Neer's evaluation, the ASES [American Shoulder and Elbow Surgeons] evaluation and the most frequent complaints of patients observed in the shoulder practice at the University of Washington" (68). Further details on how item content was selected have not been described. Item-response theory was applied later (39).

Acceptability. All item content is easy to read and understand. Missing data are rare. Low floor and ceiling effects (32,69).

Reliability. Internal reliability/consistency: Cronbach's $\alpha = 0.85$ (38).

Test–retest reliability: intraclass correlation coefficients 0.97 and 0.99 (37,69).

Validity. *Content validity.* Low floor and ceiling effects (32,69). Score distribution has not been further examined.

Criterion validity. In the absence of a gold standard, the obvious content validity of the used items and the testing studies give a certain intrinsic validity to the SST. Factor analysis revealed a 2-factor solution and questions the 1-factor total score (38). Across the entire continuum of shoulder functioning, function was not measured with equal precision but with very large confidence intervals, i.e., larger than the ASES and Shoulder Pain and Disability Index (SPADI) (39). In Rasch analysis, items 2 (. . . shoulder allows you to sleep comfortably?) and 1 (is your shoulder comfortable . . . at rest?) showed misfit (39).

Construct validity. Pearson's or Spearman's correlations of the SST to other instruments are as follows:
SPADI: 0.74 and 0.80 (ref. 32,38)
ASES: 0.73 and 0.81 (ref. 32,69)
DASH: 0.72 (ref. 71)
CS: 0.70 (ref. 72)
Western Ontario Rotator Cuff index: 0.68 (ref. 71)
SF-36 bodily pain: 0.62 (ref. 32)
SF-36 physical functioning: 0.58 (ref. 32)
SF-12 PCS: 0.44 (ref. 69)
SF-36 PCS: 0.40 and 0.60 (ref. 32,71)
SF-36 MCS: 0.16 (ref. 71)

Ability to detect change. Minimally detectable change (MDC95%) for the range 0–100: 32.3 (38).

MCID for the range 0–12: 2.05 and 2.33 for rotator cuff disease (49); 3 points for shoulder arthroplasty (70). Corresponds to MCID 17.1–25.0 for the range 0–100.

Effect sizes (ES) and standardized response means (SRMs) of the SST are as follows:
Osteoarthritis: shoulder arthroplasty: ES 2.17–2.87, SRM 1.43–1.94 (ref. 70)
Rotator cuff: repair: SRM 1.09 (ref. 71)
Injury: rotator cuff surgery: ES 1.08, SRM 1.01 (ref. 69)
Rotator cuff surgery + total shoulder arthroplasty: SRM 0.87 (ref. 37)
Injury: instability surgery: ES 0.61, SRM 0.63 (ref. 69)

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Very short and easy to use. Good construct validity.

Caveats and cautions. Substantial lack of criterion validity (testing data). Due to binary response options, questionable use of the SST score as a metric measure, especially for responsiveness (as analogously shown by versions 1 and 2 of the SF-36). In the case where an MDC95% is reported to be higher than the MCID, the MDC95% should be taken as the MCID.

Clinical usability. Easy to use; widespread use in the US. Due to lack of testing data or insufficient measurement properties, caution is necessary for measurement at an individual patient level.

Research usability. Limited due to lack of non-English versions and the caveats.

OXFORD SHOULDER SCORE (OSS)

Description

Purpose. Self-assessment of pain and function of the shoulder. Settings: shoulder operations other than stabilization (73). First published in 1996 (73). "Revision" in 2009 concerns only the specifications for use, not the content (74).

Content and number of items. 12 items: 4 about pain (2 for pain, 2 for interference with pain) and 8 about daily functions.

Response options/scale. Each item is scored into 5 Likert categories: 1 = no pain/easy to do, 2 = mild pain/little difficulty, 3 = moderate pain/moderate difficulty, 4 = severe pain/extreme difficulty, and 5 = unbearable/impossible to do. In the revision study and on the online form (see below), the item scoring is 0 (worst) to 4 (best).

Recall period for items. 4 weeks.

Endorsements. None.

Examples of use. Relevant settings (surgery; aims and analysis [references]) for the OSS are as follows:
Degenerative, inflammatory (development of the OSS, revision [73,74])
Degenerative, inflammatory (validity, responsiveness [64])
Subacromial impingement (reliability, validity, responsiveness [43,61,75])
Rotator cuff (responsiveness [35,76,77])
Osteoarthritis (responsiveness [78])
Proximal humerus fracture (validity [79])

Practical Application

How to obtain. Original published in 1 study (73) and online at http://phi.uhce.ox.ac.uk/pdf/OxfordScores/Oxshoulderscore.pdf. Online form for automatic calculation is found at http://www.orthopaedicscore.com/scorepages/oxford_shoulder_score.html.

Method of administration. Self-assessment.

Time to complete. 2 minutes.

Scoring. The (total) score is the sum of the (completed) 12 items (scoring 1–5): 12 = best and 60 = worst (73). In the revision, it is 0 = worst and 48 = best (item scoring 0–4) (74). The online form (see above) also scores on 0–48. However, missing items are scored by a 5, which is a mistake on the online form that may lead to wrong scores. How to deal with missing items has only been described for the revision: ≥10 of 12 items have to be completed (74). To compare with other instruments, we recommend total score = (m − 1) × 25, where m = mean of the completed items (originally scaled 1–5, where 5 = worst): 0 = best and 100 = worst or transformed by (100 − total score) into 0 = worst and 100 = best, as for the Short Form 36 (SF-36), and the same for the revised item scaling 0–4 (4 = best): total score = m × 25 (64).

Score interpretation. Total score, no subscores. Originally, 12 (no disability) to 60 (maximal disability). Revised OSS and online form: 0 (maximal disability) to 48 (no disability), where 0–19 = severe arthritis, 20–29 = moderate to severe arthritis, 30–39 = mild to moderate arthritis, and 40–48 = satisfactory joint function (published on the online form; see above). Normative data have not been published.

Respondent burden. All items are easy to understand and to complete and are not emotionally sensitive.

Administrative burden. Score computation is easy and needs no explanation. No training is needed to interpret the scores. Time to administer and score: ~5 minutes.

Translations/adaptations. Dutch, Italian, and German (75).

Psychometric Information

Method of development. Open interviews of outpatients and review of established questionnaires created 22 items that were longitudinally tested in several steps, resulting in the 12-item version (73). Factor analysis or item-response theory was not used.

Acceptability. All item content is short, easy to read, and understand. Missing data are rare (74). Very low floor and ceiling effects were shown (64,75).

Reliability. Internal reliability/consistency: Cronbach's $\alpha = 0.94$ (75).

Test–retest reliability: Pearson's correlation = 0.98 (75). Intraclass correlation coefficient: no published data.

Validity. *Content validity.* No published data on score distribution. Low floor and ceiling effects (64,75).

Criterion validity. In the absence of a gold standard, the obvious content validity of the used items and the moderate number of published studies examining the OSS result in a moderate intrinsic validity. Rasch and factor analysis data have not been published.

Construct validity. Pearson's or Spearman's correlations of the OSS to other instruments are as follows:
CS: 0.65–0.87 (ref. 61,64,75,79)
SPADI: 0.74 and 0.85 (ref. 43,61)
DASH: 0.79 (ref. 61)
SF-36 bodily pain: 0.64–0.76 (ref. 43,61,75)
SF-36 physical functioning: 0.57–0.68 (ref. 43,61,75)
SF-36 PCS: 0.37 (ref. 43)

Ability to detect change. Minimally detectable change (MDC95%) and minimum clinically important difference (MCID): no published data.

Effect sizes (ES) and standardized response means (SRMs) of the OSS are as follows:
Osteoarthritis and rheumatoid arthritis: hemiarthroplasty: ES 2.3 (ref. 78)
Impingement, rotator cuff: surgery: ES 1.10–1.88, SRM 1.10–1.14 (ref. 61,73,76)
Rotator cuff: decompression (± cuff repair): ES 0.97 (ref. 77)
Impingement: no treatment described: ES 0.96 (ref. 43)
Degenerative, inflammatory: surgery: ES 0.61 (ref. 64)

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Very short and responsive tool, easy to complete and to score. Specially constructed for surgical interventions. Construct validity to other measures is good. No costs to obtain.

Caveats and cautions. Data about reliability and (especially criterion) validity are rather sparse. The OSS is not often used in literature. There is only 1 important study for conservative treatment (79). Due to lack of testing data (MDC95%, MCID), caution is necessary for measurement at an individual patient level.

Clinical usability. Short tool for assessment of shoulder surgery. Easy to interpret. Due to lack of testing data or insufficient measurement properties, caution is necessary for measurement at an individual patient level.

Research usability. Validity and usability for research are rather weak. Further testing is needed.

SHOULDER DISABILITY QUESTIONNAIRE (SDQ)

Description

Purpose. Self-assessment of pain-related function of the shoulder. Settings: shoulder disorders in general (mainly soft tissue). First publication of a 22-item version in the UK (SDQ-UK) in 1994, which was not frequently used thereafter (80). Further development into the original 16-item SDQ in The Netherlands (SDQ-NL) in 1998 (81). Revision in 2000 (82).

Content and number of items. 16 items describing common situations or functions that may induce symptoms (mostly pain): "My shoulder hurts when I (do). . . ."

Response options/scale. All items are scored by "yes" = 1 or "no" = 0, and "not applicable" (missing).

Recall period for items. 24 hours.

Endorsements. None.

Examples of use. Relevant settings (aims and analysis [references]) for SDQ are as follows:
General population, primary care (development of the SDQ-UK [80])
Primary care (development, responsiveness [81])
Primary care (revision, responsiveness [82])
Shoulder instruments (comparative review [83,84])
Shoulder pain (reliability, validity [85,86])
Adhesive capsulitis (responsiveness [87,88])
Rotator cuff (responsiveness [89])
Chronic shoulder pain (responsiveness [90])

Practical Application

How to obtain. Published in 2 studies (81,82). Online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1752535/pdf/v057p00082.pdf (see Appendix).

Method of administration. Self-assessment.

Time to complete. 2 minutes.

Scoring. The (total) score is calculated by dividing the number of positively scored items (value = 1) by the total of applicable/completed items and multiplying by 100.

Score interpretation. 0 = no disability and 100 = maximal disability. A missing rule, distinct cutoffs to reflect severity, and normative data have not been published.

Respondent burden. All items are easy to understand and not emotionally sensitive.

Administrative burden. Score computation is easy. Time to administer and score: ~5 minutes.

Translations/adaptations. English (80–82), Dutch (original, not published), Spanish, and Turkish (86).

Psychometric Information

Method of development. Questions considered relevant to the shoulder were selected from the Functional Limitations Profile and a list of activities from therapists and patients was added (80–82). Reduction from 60 and 22 to 16 items according to the "judgmental approach" (81,82). Data about factor analysis or item-response theory have not been published.

Acceptability. Easy to read and understand. Missing data are rare (83). A substantial ceiling effect was shown (82).

Reliability. Internal reliability/consistency: Cronbach's $\alpha = 0.76$ and 0.79, respectively (85,86).

Test–retest reliability: Pearson's correlation = 0.88 (86). Intraclass correlation coefficient: no data published.

Validity. *Content validity.* The content validity of the SDQ-NL was rated as doubtful in a comparison of multiple shoulder tools (84). There was a substantial ceiling effect (82,84).

Criterion validity. There are only sparse data on criterion validity (84). Rasch and factor analysis data have not been published.

Construct validity. Pearson's or Spearman's correlations of the SDQ to other instruments are as follows:

VAS for function: 0.58 (ref. 85)
SDQ-UK: 0.55 (ref. 83)
VAS for pain: 0.41 (ref. 85)
SPADI: 0.33 (ref. 83)
ROM: 0.27–0.41 (ref. 85)

Ability to detect change. Minimally detectable change (MDC95%) and minimum clinically important difference (MCID): no data published. A mean change score of 40 points was highly specific for improvement (81).

Effect sizes (ES) and standardized response means (SRMs) of the SDQ are as follows:

Adhesive capsulitis: mobilization: ES 5.43 and 2.81, SRM 3.88 and 3.40 (ref. 87)

Rotator cuff tendinitis: steroids, transcutaneous electrical nerve stimulation: ES 5.19 and 5.43, SRM 5.83 and 4.08 (ref. 89)

Primary care: soft tissue, physiotherapy: SRM 2.22 and 1.14 (Guyatt's responsiveness index [ref. 81,82])

Adhesive capsulitis: steroids, physiotherapy: ES 1.73 and 1.12, SRM 1.32 and 0.97 (ref. 88)

Chronic shoulder pain: graded exercise, usual care: ES 0.94 and 0.77, SRM 0.65 and 0.71 (ref. 90)

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Very short tool, easy to complete and to score. No costs to obtain.

Caveats and cautions. Data about reliability and validity are sparse. Content and criterion validity have been rated to be doubtful (84). Construct validity compared to other measures is weak. Published data indicate low reliability and validity, especially for measurement at an individual patient level (no data on MDC95%, MCID). Due to binary response options, questionable use of the SDQ score as a metric measure, especially for responsiveness (as analogously shown by versions 1 and 2 of the Short Form 36). Responsiveness results are extraordinarily high. The SDQ is rarely reported in the literature. There are no data on diseases related to the shoulder joint itself.

Clinical usability. Although the SDQ is a short and easy to interpret tool, caution is necessary for clinical usability and measurement at a group or an individual patient level due to the lack of testing data and insufficient measurement properties (see above).

Research usability. Weak and doubtful validity and usability for research. Further psychometric testing is needed.

WESTERN ONTARIO SHOULDER INSTABILITY INDEX (WOSI)

Description

Purpose. "To evaluate the disease-specific quality-of-life of patients with symptomatic shoulder instability" (67). Settings: shoulder instability. First published in 1998 (67). Revision of scaling in 2005 (91).

Content and number of items. 21 items in 4 domains: physical symptoms, including pain (10 items), sports/recreation/work (4 items), lifestyle (4 items), and emotions (3 items).

Response options/scale. Each item is scored on a 100-mm visual analog scale (VAS). The use of a corresponding 11-point numerical rating scale (NRS; 0–10) to scan forms was later approved by the developer (92).

Recall period for items. 1 week.

Endorsements. Recommended by the European Society for Surgery of the Shoulder and the Elbow (online at http://www.secec.org).

Examples of use. Relevant settings (shoulder instability; aims and analysis [references]) for the WOSI are as follows:

Various reasons for instability (development, reliability, validity, responsiveness [67])

Randomized controlled trial: rehabilitation, surgery (revision of scaling, outcome [91])

Shoulder instruments (comparative review [93,94])

Physiotherapy, surgery (German WOSI: reliability, validity, responsiveness [92,95])

Stabilization surgery (Swedish WOSI: reliability, validity, responsiveness [96])

There are several other studies that have used the WOSI, but they only report followup without baseline scores and have been excluded from the review for this reason.

Practical Application

How to obtain. Published in (67,91,92,95,96). Free online at http://www.secec.org/data/upload/files/Western%20Ontario%20Shoulder%20Instability%20Index%20(WOSI).pdf.

Method of administration. Self-assessment.

Time to complete. No data published. Estimated: 3 minutes.

Scoring. Sum of 21 unweighted items (0 = best and 100 = worst).

Score interpretation. 0 = best to 2,100 = worst. We recommend a transformed score by 100 − original score/21 ranging from 0 = worst to 100 = best to be comparable to other instruments, as the Short Form 36 (SF-36) (91). A missing rule (we recommend at least 2 of 3 = 14 of 21 completed items), distinct cutoffs for severity, and normative data have not been published.

Respondent burden. Minimal; easy questions and use of the VAS.

Administrative burden. Reduced with the use of the 11-point NRS. Easy computation of the total score. Estimated 6 minutes.

Translations/adaptations. Validated versions are available in Swedish (96) and in German: 2 simultaneously published versions, one is approved by the developer (92) and the other is not approved (95).

Psychometric Information

Method of development. Item generation by review of the literature (other instruments) and interview of specialists and patients with shoulder instability (67). Item reduction based on expert group, empirical testing (patient's perception of item importance), and inter-item correlation (67).

Acceptability. Highly accepted by patients and surgeons because of the importance of the item contents. No floor and ceiling effects (92,96). The WOSI got the best rating of psychometric properties in a systematic review (93).

Reliability. Internal reliability/consistency: Cronbach's $\alpha = 0.88-0.96$ (92,95,96).

Test–retest reliability: intraclass correlation coefficient 0.87–0.98 (67,92,95,96).

Validity. *Content validity.* Item content established by patients and experts. No floor or ceiling effects (92,96). Score distribution has not been further examined.

Criterion validity. There is no gold standard to measure shoulder instability. The obvious content validity of the items and the data from the psychometric testing studies result in a certain intrinsic validity. No data on item-response theory, factor, or Rasch analysis have been published.

Construct validity. Pearson's or Spearman's correlations of the WOSI to other instruments are as follows:

VAS for function:	0.80 (ref. 96)
DASH:	0.77 (ref. 67)
SF-12 PCS:	0.66 (ref. 67)
CS:	0.59 (ref. 95)
Rowe score:	0.59 (ref. 96)
Shoulder rating scale:	0.59 (ref. 67)
SF-36 bodily pain:	0.56 (ref. 95)
ASES:	0.55–0.67 (ref. 67,92)
SF-36 physical functioning:	0.44 (ref. 95)
EQ-5D:	0.44 (ref. 96)
SF-12 MCS:	0.12 (ref. 67)

Ability to detect change. Minimally detectable change (MDC95%) and minimum clinically important difference (MCID): no data published.

Effect sizes (ES) and standardized response means (SRMs) of the WOSI are as follows:

Stabilization surgery: ES 1.67, SRM 1.40 (ref. 96)
Physiotherapy ± stabilization surgery: SRM 0.93 (ref. 67)

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Relevant questions and domains, high patient acceptance, good construct validity. Psychometrically best-tested tool for shoulder instability (96).

Caveats and cautions. Substantial lack of validity and responsiveness testing data. Due to lack of testing data (MDC95%, MCID), caution is necessary for measurement at an individual patient level. Unusual scale from 0 = worst to 2,100 = best in the original scaling.

Clinical usability. Highly accepted by patients. Easy to use as patient self-assessment, no clinical examination necessary (as often to be done for instability). Due to lack of testing data or insufficient measurement properties, caution is necessary for measurement at an individual patient level.

Research usability. Has to be recommended as the best psychometrically tested tool for shoulder instability, also in a set of different instruments if instability is present (96). However, there is still lack of testing data.

ACKNOWLEDGMENTS

We gratefully thank Susanne Lehmann and Franziska Kohler for their work in procuring and organizing the literature, Carol Kennedy and Dorcas E. Beaton for commenting and editing the DASH chapter, Kathreen E. Roach for commenting and editing the SPADI chapter, Susann Drerup for commenting and editing the WOSI chapter, and Joy Buchanan for her English editing.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Kennedy CA, Beaton DE, Solway S, McConnell S, Bombardier C. The DASH outcome measure user's manual. 3rd ed. Toronto: Institute for Work & Health; 2011.
2. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (Disabilities of the Arm, Shoulder and Hand). Am J Ind Med 1996;29:602–8.
3. Beaton DE, Wright JG, Katz JN, and the Upper Extremity Collaborative Group. Development of the *Quick*DASH: comparison of three item-reduction approaches. J Bone Joint Surg Am 2005;87:1038–46.
4. Gabel CP, Yelland M, Melloh M, Burkett B. A modified QuickDASH-9 provides a valid outcome instrument for upper limb function. BMC Musculoskelet Disord 2009;10:161.
5. Germann G, Harth A, Wind G. Standardisation and validation of the German version 2.0 of the Disability of Arm, Shoulder, Hand (DASH) questionnaire. Unfallchirurg 2003;105:13–9. In German.
6. Angst F, Pap G, Mannion AF, Herren DB, Aeschlimann A, Schwyzer HK, et al. Comprehensive assessment of clinical outcome and quality of life after total shoulder arthroplasty: usefulness and validity of subjective outcome measures. Arthritis Rheum 2004;51:819–28.
7. Roy JS, MacDermid JC, Woodhouse LJ. Measuring shoulder function: a systematic review of four questionnaires. Arthritis Rheum 2009;61: 623–32.
8. Cano S, Barrett L, Zajicek J, Hobart JC. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. Mult Scler 2011;17:214–22.
9. Franchignoni F, Giordano A, Sartorio F, Vercelli S, Pascariello B, Ferriero G. Suggestions for refinement of the Disabilities of the Arm, Shoulder and Hand outcome measure (DASH): a factor analysis and Rasch validation study. Arch Phys Med Rehabil 2010;91:1370–7.
10. Gummesson C, Ward MM, Atroshi I. The shortened Disabilities of the

Arm, Shoulder and Hand questionnaire (QuickDASH): validity and reliability based on responses within the full-length DASH. BMC Musculoskelet Disord 2006;7:44.

11. Angst F, Goldhahn J, Drerup S, Flury M, Schwyzer HK, Simmen BR. How sharp is the short QuickDASH? A refined content and validity analysis of the short form of the Disabilities of the Shoulder, Arm and Hand questionnaire in the strata of symptoms and function and specific joint conditions. Qual Life Res 2009;18:1043–51.
12. Hunsaker FG, Cioffi DA, Amadio PC, Wright JG, Caughlin B. The American Academy of Orthopaedic Surgeons outcomes instruments: normative values from the general population. J Bone Joint Surg Am 2002;84:208–15.
13. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis 2004;63:335–41.
14. Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V. Measuring the whole or parts? Validity, reliability, and responsiveness of the Disabilities of the Arm Shoulder and Hand outcome measure in different regions of the upper extremity. J Hand Ther 2001;14:128–46.
15. Gummesson C, Atroshi I, Ekdahl C. The Disabilities of the Arm, Shoulder and Hand (DASH) outcome questionnaire: longitudinal construct validity and measuring self-rated health change after surgery. BMC Musculoskelet Disord 2003;4:11.
16. Huisstede BM, Feleus A, Bierma-Zeinstra SM, Verhaar JA, Koes BW. Is the Disability of Arm, Shoulder, and Hand questionnaire (DASH) also valid and responsive in patients with neck complaints? Spine (Phila Pa 1976) 2009;34:E130–8.
17. Dowrick AS, Gabbe BJ, Williamson OD, Cameron PA. Does the Disabilities of the Arm, Shoulder and Hand (DASH) scoring system only measure disability due to injuries to the upper limb? J Bone Joint Surg Br 2006;88:524–7.
18. Raven EE, Haverkamp D, Sierevelt IN, van Montfoort DO, Poll RG, Blankevoort L, et al. Construct validity and reliability of the Disability of Arm, Shoulder and Hand questionnaire for upper extremity complaints in rheumatoid arthritis. J Rheumatol 2008;35:2334–8.
19. Angst F, Goldhahn J, Drerup S, Aeschlimann A, Schwyzer HK, Simmen BR. Responsiveness of six outcome assessment instruments in total shoulder arthroplasty. Arthritis Rheum 2008;59:391–8.
20. Staples MP, Forbes A, Green S, Buchbinder R. Shoulder-specific disability measures showed acceptable construct validity and responsiveness. J Clin Epidemiol 2010;63:163–70.
21. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. J Clin Epidemiol 2004;57:1008–18.
22. Slobogean GP, Noonan VK, O'Brien PJ. The reliability and validity of the Disabilities of Arm, Shoulder, and Hand, EuroQol-5D, Health Utilities Index, and Short Form-6D outcome instruments in patients with proximal humeral fractures. J Shoulder Elbow Surg 2010;19:342–8.
23. Turchin DC, Beaton DE, Richards RR. Validity of observer-based aggregate scoring systems as descriptors of elbow pain, function, and disability. J Bone Joint Surg Am 1998;80:154–62.
24. Angst F, John M, Pap G, Mannion AF, Herren DB, Flury M, et al. Comprehensive assessment of clinical outcome and quality of life after total elbow arthroplasty. Arthritis Rheum 2005;53:73–82.
25. MacDermid JC, Richards RS, Donner A, Bellamy N, Roth JH. Responsiveness of the Short Form-36, Disability of the Arm, Shoulder, and Hand questionnaire, patient-rated wrist evaluation, and physical impairment measurements in evaluating recovery after a distal radius fracture. J Hand Surg Am 2000;25:330–40.
26. MacDermid JC, Tottenham V. Responsiveness of the Disability of the Arm, Shoulder, and Hand (DASH) and Patient-Rated Wrist/Hand Evaluation (PRWHE) in evaluating change after hand therapy. J Hand Ther 2004;17:18–23.
27. Angst F, John M, Goldhahn J, Herren DB, Pap G, Aeschlimann A, et al. Comprehensive assessment of clinical outcome and quality of life after resection interposition arthroplasty of the thumb saddle joint. Arthritis Rheum 2005;53:205–13.
28. Mintken PE, Glynn P, Cleland JA. Psychometric properties of the shortened Disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and numeric pain rating scale in patients with shoulder pain. J Shoulder Elbow Surg 2009;18:920–6.
29. Beaton DE, Eerd D, Smith P, van der Velde G, Cullen K, Kennedy CA, et al. Minimal change is sensitive, less specific, to recovery: a diagnostic testing approach to interpretability. J Clin Epidemiol 2011;64:487–96.
30. Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a Shoulder Pain and Disability Index. Arthritis Care Res 1991;4:143–9.
31. Angst F, Goldhahn J, Pap G, Mannion AF, Roach KE, Siebertz D, et al. Cross-cultural adaptation, reliability, and validity of the German Shoulder Pain and Disability Index (SPADI). Rheumatology (Oxford) 2007;46:87–92.
32. Beaton DE, Richards RR. Measuring function of the shoulder: a cross-sectional comparison of five questionnaires. J Bone Joint Surg Am 1996;78:882–90.
33. Tveita EK, Sanvik L, Ekeberg OM, Juel NG, Bautz-Holter E. Factor structure of the Shoulder Pain and Disability Index in patients with adhesive capsulitis. BMC Musculoskelet Disord 2008;9:103.
34. Tveita EK, Ekeberg OM, Juel NG, Bautz-Holter E. Responsiveness of the Shoulder Pain and Disability Index in patients with adhesive capsulitis. BMC Musculoskelet Disord 2008;9:161.
35. Ekeberg OM, Bautz-Holter E, Tveita EK, Keller A, Juel NG, Brox JI. Agreement, reliability and validity in 3 shoulder questionnaires in patients with rotator cuff disease. BMC Musculoskelet Disord 2008;9:68.
36. Ekeberg OM, Bautz-Holter E, Keller A, Tveita EK, Juel NG, Brox JI. A questionnaire found disease-specific WORC index is not more responsive than SPADI and OSS in rotator cuff disease. J Clin Epidemiol 2010;63:575–84.
37. Beaton DE, Richards RR. Assessing the reliability and responsiveness of 5 shoulder questionnaires. J Shoulder Elbow Surg 1998;7:565–72.
38. Roddey TS, Olson SL, Cook KF, Gartsman GM, Hanten W. Comparison of the University of California-Los Angeles Shoulder Scale and the Simple Shoulder Test with the Shoulder Pain and Disability Index: single-administration reliability and validity. Phys Ther 2000;80:759–68.
39. Cook KF, Gartsman GM, Roddey TS, Olson SL. The measurement level and trait-specific reliability of 4 scales of shoulder functioning: an empiric investigation. Arch Phys Med Rehabil 2001;82:1558–65.
40. Williams JW Jr, Holleman DR Jr, Simel DL. Measuring shoulder function with the Shoulder Pain and Disability Index. J Rheumatol 1995;22:727–32.
41. Heald SL, Riddle DL, Lamb RL. The Shoulder Pain and Disability Index: the construct validity and responsiveness of a region-specific disability measure. Phys Ther 1997;77:1079–89.
42. MacDermid JC, Solomon P, Prkachin K. The Shoulder Pain and Disability Index demonstrates factor, construct and longitudinal validity. BMC Musculoskelet Disord 2006;7:12.
43. Cloke DJ, Lynn SE, Watson H, Steen IN, Purdy S, Williams JR. A comparison of functional, patient-based scores in subacromial impingement. J Shoulder Elbow Surg 2005;14:380–4.
44. Richards RR, An KN, Bigliani LU, Friedman RJ, Gartsman GM, Gristina AG, et al. A standardized method for the assessment of shoulder function. J Shoulder Elbow Surg 1994;3:347–52.
45. Sallay PI, Reed L. The measurement of normative American Shoulder and Elbow Surgeons scores. J Shoulder Elbow Surg 2003;12:622–7.
46. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons standardized shoulder assessment form, patient self-report section: reliability, validity, and responsiveness. J Shoulder Elbow Surg 2002;11:587–94.
47. Razmjou H, Bean A, van Osnabrugge V, MacDermid JC, Holtby R. Cross-sectional and longitudinal construct validity of two rotator cuff disease-specific outcome measures. BMC Musculoskelet Disord 2006;7:26.
48. Yoo JC, Koh KH, Park WH, Park JC, Kim SM, Yoon YC. The outcome of ultrasound-guided needle decompression and steroid injection in calcific tendinitis. J Shoulder Elbow Surg 2010;19:596–600.
49. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. J Bone Joint Surg Am 2010;92:296–303.
50. Padua R, Padua L, Ceccarelli E, Bondi R, Alviti F, Castagna A. Italian version of ASES questionnaire for shoulder assessment: cross-cultural adaptation and validation. Musculoskelet Surg 2010;94 Suppl:S85–90.
51. Kocher MS, Horan MP, Briggs KK, Richardson TR, O'Holleran J, Hawkins RJ. Reliability, validity, and responsiveness of the American Shoulder and Elbow Surgeons subjective shoulder scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. J Bone Joint Surg Am 2005;87:2006–11.
52. Goldhahn J, Angst F, Drerup S, Pap G, Simmen BR, Mannion AF. Lessons learned during the cross-cultural adaptation of the American Shoulder and Elbow Surgeons shoulder form into German. J Shoulder Elbow Surg 2008;17:248–54.
53. Cook KF, Roddey TS, Olson SL, Gartsman GM, Valenzuela FF, Hanten WP. Reliability by surgical status of self-reported outcomes in patients who have shoulder pathologies. J Orthop Sports Phys Ther 2002;32:336–46.
54. Lo IK, Litchfield RB, Griffin S, Faber K, Patterson SD, Kirkley A. Quality-of-life outcome following hemiarthroplasty or total shoulder arthroplasty in patients with osteoarthritis: a prospective, randomized trial. J Bone Joint Surg Am 2005;87:2178–85.
55. Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. Clin Orthop Relat Res 1987;214:160–4.
56. Constant CR, Gerber C, Emery RJ, Sojbjerg JO, Gohlke F, Boileau P. A

review of the Constant score: modifications and guidelines for its use. J Shoulder Elbow Surg 2008;17:355–61.

57. Roy JS, MacDermid JC, Woodhouse LJ. A systematic review of the psychometric properties of the Constant-Murley score. J Shoulder Elbow Surg 2010;19:157–64.
58. Yian EH, Ramappa AJ, Arneberg O, Gerber C. The Constant Score in normal shoulders. J Shoulder Elbow Surg 2005;14:128–33.
59. Rocourt MH, Radlinger L, Kalberer F, Sanavi S, Schmid NS, Leunig M, et al. Evaluation of intratester and intertester reliability of the Constant-Murley shoulder assessment. J Shoulder Elbow Surg 2008;17:364–9.
60. Oh JH, Jo KH, Kim WS, Gong HS, Han SG, Kim YH. Comparative evaluation of the measurement properties of various shoulder outcome instruments. Am J Sports Med 2009;37:1161–8.
61. Dawson J, Hill G, Fitzpatrick R, Carr A. The benefits of using patient-based methods of assessment: medium-term results of an observational study of shoulder surgery. J Bone Joint Surg Br 2001;83:877–82.
62. Vas J, Ortega C, Olmo V, Perez-Fernandez F, Hernandez L, Medina I, et al. Single-point acupuncture and physiotherapy for the treatment of painful shoulder: a multicentre randomized controlled trial. Rheumatology (Oxford) 2008;47:887–93.
63. O'Connor DA, Chipchase LS, Tomlinson J, Krishnan J. Arthroscopic subacromial decompression: responsiveness of disease-specific and health-related quality of life outcome measures. Arthroscopy 1999;15: 836–40.
64. Christie A, Hagen KB, Mowinckel P, Dagfinrud H. Methodological properties of six shoulder disability measures in patients with rheumatic diseases referred for shoulder surgery. J Shoulder Elbow Surg 2009;18:89–95.
65. Skutek M, Fremerey RW, Zeichen J, Bosch U. Outcome analysis following open rotator cuff repair: early effectiveness validated using four different shoulder assessment scales. Arch Orthop Trauma Surg 2000; 120:432–6.
66. Razmjou H, Bean A, Macdermid JC, van Osnabrugge V, Travers N, Holtby R. Convergent validity of the Constant-Murley outcome measure in patients with rotator cuff disease. Physiother Can 2008;60:72–9.
67. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability: the Western Ontario Shoulder Instability Index (WOSI). Am J Sports Med 1998;26:764–72.
68. Lippitt SB, Harryman DT, Matsen FA. A practical tool for evaluation of function: the Simple Shoulder Test. In: Matsen FA, Fu FH, Hawkins RJ, editors. The shoulder: a balance of mobility and stability. Rosemont (IL): American Academy of Orthopedic Surgeons; 1993. p. 545–59.
69. Godfrey J, Hamman R, Lowenstein S, Briggs K, Kocher M. Reliability, validity, and responsiveness of the Simple Shoulder Test: psychometric properties by age and injury type. J Shoulder Elbow Surg 2007;16: 260–7.
70. Roy JS, Macdermid JC, Faber KJ, Drosdowech DS, Athwal GS. The Simple Shoulder Test is responsive in assessing change following shoulder arthroplasty. J Orthop Sports Phys Ther 2010;40:413–21.
71. MacDermid JC, Drosdowech D, Faber K. Responsiveness of self-report scales in patients recovering from rotator cuff surgery. J Shoulder Elbow Surg 2006;15:407–14.
72. Romeo AA, Mazzocca A, Hang DW, Shott S, Bach BR. Shoulder scoring scales for the evaluation of rotator cuff repair. Clin Orthop Rel Res 2004;427:107–14.
73. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. J Bone Joint Surg Br 1996;78:593–600.
74. Dawson J, Rogers K, Fitzpatrick R, Carr A. The Oxford Shoulder Score revisited. Arch Orthop Trauma Surg 2009;129:119–23.
75. Huber W, Hofstaetter JG, Hanslik-Schnabel B, Posch M, Wurnig C. The German version of the Oxford Shoulder Score: cross-cultural adaptation and validation. Arch Orthop Trauma Surg 2004;124:531–6.
76. Dawson J, Hill G, Fitzpatrick R, Carr A. Comparison of clinical and patient-based measures to assess medium-term outcomes following shoulder surgery for disorders of the rotator cuff. Arthritis Rheum 2002;47:513–9.
77. Allom R, Colegate-Stone T, Gee M, Ismail M, Sinha J. Outcome analysis of surgery for disorders of the rotator cuff: a comparison of subjective and objective scoring tools. J Bone Joint Surg Br 2009;91:367–73.
78. Rees JL, Dawson J, Hand GC, Cooper C, Judge A, Price AJ, et al. The use of patient-reported outcome measures and patient satisfaction ratings to assess outcome in hemiarthroplasty of the shoulder. J Bone Joint Surg Br 2010;92:1107–11.
79. Baker P, Nanda R, Goodchild L, Finn P, Rangan A. A comparison of the Constant and Oxford Shoulder Scores in patients with conservatively treated proximal humeral fractures. J Shoulder Elbow Surg 2008;17: 37–41.
80. Croft P, Pope D, Zonca M, O'Neill T, Silman A. Measurement of shoulder disability: results of a validation study. Ann Rheum Dis 1994;53:525–8.
81. Van der Windt DA, van der Heijden GJ, de Winter AF, Koes BW, Deville W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. Ann Rheum Dis 1998;57:82–7.
82. Van der Heijden GJ, Leffers P, Bouter LM. Shoulder Disability Questionnaire design and responsiveness of a functional status measure. J Clin Epidemiol 2000;53:29–38.
83. Paul A, Lewis M, Shadforth MF, Croft PR, van der Windt DA, Hay EM. A comparison of four shoulder-specific questionnaires in primary care. Ann Rheum Dis 2004;63:1293–9.
84. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Decker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis 2004;58:530–40.
85. De Winter A, van der Heijden GJ, Scholten RJ, van der Windt DA, Bouter LM. The Shoulder Disability Questionnaire differentiated well between high and low disability levels in patients in primary care, in a cross-sectional study. J Clin Epidemiol 2007;60:1156–63.
86. Ozsahin M, Akgun K, Aktas I, Kurtais Y. Adaptation of the Shoulder Disability Questionnaire to the Turkish population, its reliability and validity. Int J Rehabil Res 2008;31:241–5.
87. Vermeulen HM, Rozing PM, Obermann WR, le Cessie S, Vlieland TP. Comparison of high-grade and low-grade mobilization techniques in the management of adhesive capsulitis of the shoulder: randomised controlled study. Phys Ther 2006;86:355–68.
88. Ryans I, Montgomery A, Galway R, Kernohan WG, McKane R. A randomized controlled trial of intra-articular triamcinolone and/or physiotherapy in shoulder capsulitis. Rheumatology (Oxford) 2005;44: 529–35.
89. Eyigor C, Eyigor S, Korkmaz OK. Are intra-articular corticosteroid injections better than conventional TENS in treatment of rotator cuff tendinitis in the short run? A randomized study. Eur J Phys Rehabil Med 2010;46:315–24.
90. Geraets JJ, Goossens ME, de Groot IJ, de Bruijn CP, de Bie RA, Dinant GJ, et al. Effectiveness of a graded exercise program for patients with chronic shoulder complaints. Austr J Physiother 2005;51:87–94.
91. Kirkley A, Werstine R, Ratjek A, Griffin S. Prospective randomized clinical trial comparing the effectiveness of immediate arthroscopic stabilization versus immobilization and rehabilitation in first traumatic anterior dislocations of the shoulder: long-term evaluation. Arthroscopy 2005;21:55–63.
92. Drerup S, Angst F, Griffin S, Flury MP, Simmen BR, Goldhahn J. Western Ontario shoulder instability index (WOSI): translation and cross-cultural adaptation for use by German speakers. Orthopade 2010; 39:711–8. In German.
93. Rouleau DM, Faber K, MacDermid JC. Systematic review of patient-administered shoulder functional scores on instability. J Shoulder Elbow Surg 2010;19:1121–8.
94. Kirkley A, Griffin S, Dainty K. Scoring systems for the functional assessment of the shoulder. Arthroscopy 2003;19:1109–20.
95. Hofstaetter JG, Hanslik-Schnabel B, Hofstaetter SG, Wurnig C, Huber W. Cross-cultural adaptation and validation of the German version of the Western Ontario Shoulder Instability index. Arch Orthop Trauma Surg 2010;130:787–96.
96. Salomonsson B, Ahlstrom S, Dalen N, Lillkrona U. The Western Ontario Shoulder Instability Index (WOSI): validity, reliability, and responsiveness retested with a Swedish translation. Acta Orthop 2009; 80:233–8.

Summary Table for Adult Shoulder Function Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
DASH	Symptoms and function of the entire upper extremity 30 items	Self-assessment	4 minutes Easy to understand	10 minutes Hand or computer score	Total score and subscores symptoms and function 0 = best, 100 = worst	Cronbach's α = 0.92–0.98 ICC 0.93–0.98	Content and construct very high Criterion moderate	Moderately responsive for shoulder conditions	Best tested, most widely used Also for multilocular conditions C+, R+	Long, strict missing rule, not shoulder specific
*Quick*DASH	Extraction of the DASH: symptoms and function of the entire upper extremity 11 items	Self-assessment	2 minutes Easy to understand	8 minutes Hand or computer score	Total score 0 = best, 100 = worst	Cronbach's α = 0.92–0.95 ICC 0.90–0.94	Content and construct high Criterion moderate to high	Moderately responsive for shoulder conditions	Very well tested, very short Also for multilocular conditions C+, R(+)	Total score only. Not exactly the same construct as the DASH, not shoulder specific
SPADI	Pain and function 13 items	Self-assessment	2 minutes Easy to understand	5 minutes Hand or computer score	Total score and subscores pain and function 0 = best, 100 = worst	Cronbach's α = 0.86–0.96 ICC 0.84–0.95	Content and construct high Criterion moderate to high	Highly responsive	Very well tested, short, responsive C+, R+	Few proviso regarding validity
ASES	pASES: pain, ADL, instability 16 items cASES: range of motion, signs, strength, instability 34 items	Self- and examiner assessment	pASES: 3 minutes Easy to understand cASES: clinical examination is time consuming	8 minutes (pASES) Hand or computer score	Total score and subscores 0 = worst, 100 = best.	Cronbach's α = 0.61–0.96 ICC 0.84–0.96	Content and construct moderate to high Criterion: questionable but lack of data	Highly responsive	Moderately tested, responsive, widespread use C+, R(+)	Long tool, not easy to score and interpret. Mix of different item scales. Limited criterion validity. Sparse data for the cASES
CS	Pain, ADL, mobility, strength 10 items	Interviewer (self-) and examiner assessment combined	5–7 minutes Easy to understand	10 minutes Hand score	Total score 0 = worst, 100 = best (subscores' maximum = best: 15, 20, 40, 25)	Cronbach's α = 0.37 and 0.60 ICC 0.80–0.96 Low intertester reliability	Content moderate Criterion low Construct high	Highly responsive	Most used (partly) clinical tool, widely accepted, responsive C+, R(–)	Lack in reliability and validity (testing). Different protocols for measuring strength
SST	Function 12 items	Self-assessment	2–3 minutes Very easy to understand	5 minutes Hand score	Total score 0 = worst, 100 (originally 12) = best	Cronbach's α = 0.85 ICC 0.97 and 0.99	Content and construct moderate to high Criterion low	Doubtfully/ moderately responsive	Widely used in US, very short and simple C(+), R–	Lack in criterion validity. Questionable use as metric score
OSS	Pain and function 12 items	Self-assessment	2 minutes Easy to understand	5 minutes Hand or computer score	Total score 0 = worst, 48 = best	Cronbach's α = 0.94 ICC: no data	Content moderate Construct high Criterion: lack of data	Highly responsive	Very short, specific for surgery C(+), R–	Lack in reliability and validity testing, especially in nonsurgical conditions
SDQ	Pain-related function 16 items	Self-assessment	2 minutes Very easy to understand	5 minutes Hand score	Total score 0 = best, 100 = worst	Cronbach's α = 0.76 and 0.79 ICC: no data	Low/questionable validity in all domains	Doubtfully responsive	Very short, easy to determine C–, R–	Lack in reliability and validity (testing). Questionable use as metric score
WOSI	(In)stability 21 items	Self-assessment	3 minutes Easy to understand	6 minutes Hand or computer score	Total score 0 = best, 2,100 = worst	Cronbach's α = 0.88–0.96 ICC 0.87–0.98	Content and construct moderate to high Criterion: sparse data	Well responsive but lack of data	Best-tested tool for instability, high clinical acceptance C(+), R(+)	Still lack of data about validity and responsiveness

* DASH = Disabilities of the Arm, Shoulder, and Hand questionnaire; ICC = intraclass correlation coefficient; C = clinical use; + = appropriate/to be recommended; R = use for research; *Quick*DASH = short form of the DASH; parentheses = use with caution; SPADI = Shoulder Pain and Disability Index; ASES = American Shoulder and Elbow Surgeons questionnaire for the shoulder; pASES = patient ASES; ADL = activities of daily living; cASES = clinical ASES; CS = Constant (Murley) Score; – = not appropriate/not to be recommended; SST = Simple Shoulder Test; OSS = Oxford Shoulder Score; SDQ = Shoulder Disability Questionnaire; WOSI = Western Ontario Shoulder Instability index.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S350–S370
DOI 10.1002/acr.20538
© 2011, American College of Rheumatology

Measures of Physical Performance Assessments

Self-Paced Walk Test (SPWT), Stair Climb Test (SCT), Six-Minute Walk Test (6MWT), Chair Stand Test (CST), Timed Up & Go (TUG), Sock Test, Lift and Carry Test (LCT), and Car Task

KIM BENNELL, FIONA DOBSON, AND RANA HINMAN

INTRODUCTION

In this review, clinical physical performance measures (PPMs) that relate directly to people with lower extremity osteoarthritis (OA) (1), yet are also relevant for other rheumatic conditions that affect the lower extremity, are evaluated. This information is complementary and an update to some of the measures of adult general performance presented in the special issue of *Arthritis Care & Research* in 2003 (2). In the current review, PPMs are defined as clinician-observed measures of physical function that assess a task that can be classified as "activities" using the World Health Organization International Classification of Functioning, Disability and Health (ICF) model (3). They do not include measures that are specific tests of body structure, body function, or impairments, e.g., specific measures of strength or balance. Physical function is related to "the ability to move around" (4) and "the ability to perform daily activities" (5) and is assessed directly by an observer while the activity is being performed by an individual, usually by timing, counting, or distance measures. PPMs measure what an individual can do rather than what the individual perceives they can do as in self-reported functional measures (5).

The selection of PPMs for this review was based on the following criteria: 1) clinical (field) tests: PPMs were selected if they were readily available, required portable or no equipment, and could be conducted within the clinical setting; 2) relevant to core activities commonly impaired in people with OA: a range of PPMs was selected to reflect the ICF activities most relevant to individuals with lower extremity OA, including walking and moving (ICF d450–69), changing and maintaining body position (ICF d410–29), climbing (ICF d4551), and carrying, moving, and handling objects (ICF d430–49); 3) current trends: PPMs that have been included in a performance battery for lower extremity OA were targeted to reflect current trends and recommendations in recent literature (a performance battery is a composite of a number of individual PPMs grouped together); and 4) most commonly cited: PPMs for individuals with OA that were most commonly cited in a literature search were given priority over those less frequently cited.

A computerized literature search using Medline, CINAHL, ISI Web of Science, Scopus, and Cochrane was performed. Key terms were mapped to medical subject headings terms: osteoarthritis (hip and knee), task performance and analysis, observation, physical examination, walking or mobility limitations, physical fitness, physical functioning or disability evaluation, and performance-based measures.

Kim Bennell, PhD, Fiona Dobson, PhD, Rana Hinman, PhD: University of Melbourne, Melbourne, Victoria, Australia.

Address correspondence to Kim Bennell, PhD, University of Melbourne, Centre for Health, Exercise and Sports Medicine, 200 Berkeley Street, Parkville, Melbourne, Victoria, Australia 3010. E-mail: k.bennell@unimelb.edu.au.

Submitted for publication December 21, 2010; accepted in revised form June 16, 2011.

SELF-PACED WALK TEST (SPWT)

Description

Purpose. The SPWT assesses the time it takes to walk short distances (typically less than 50 meters/150 feet). A number of different distances have been reported for the SPWT, e.g., 8 feet (6), 13 meters (7), 50 feet (8–10), or 40 meters (11–13). The SPWT is used in many population groups, including hip and knee osteoarthritis (OA), rheumatoid arthritis (RA), older adults, and children.

Content. Individuals are asked to walk quickly and safely without overexerting themselves. The time it takes to cover a specified distance is recorded in seconds.

Domains covered. Walking short distances.

International Classification of Functioning, Disability and Health categories. d450: walking, d410–d429: changing and maintaining body position.

Number of items. 1.

Response options/scale. Time (seconds). Measured on a continuous ratio scale.

Recall period for items. Not applicable (N/A).

Endorsements. Cibulka MT, White DM, Woehrle J, Harris-Hayes M, Enseki K, Fagerson TL, et al. Hip pain and mobility deficits. Hip osteoarthritis: clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopaedic section of the American Physical Therapy Association. J Orthop Sports Phys Ther 2009;39:A1–25 (14).

Examples of use. Outcome measure following hip/knee arthroplasty: Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskelet Disord 2005;6:3 (11).

Outcome measure following physical rehabilitation (exercise programs) for hip and knee OA: Galea MP, Levinger P, Lythgo N, Cimoli C, Weller R, Tully E, et al. A targeted home- and center-based exercise program for people after total hip replacement: a randomized clinical trial. Arch Phys Med Rehabil 2008;89:1442–7 (15).

Silva LE, Valim V, Pessanha AP, Oliveira LM, Myamoto S, Jones A, et al. Hydrotherapy versus conventional land-based exercise for the management of patients with osteoarthritis of the knee: a randomized clinical trial. Phys Ther 2008;88:12–21 (16).

Fisher NM, Gresham GE, Abrams M, Hicks J, Horrigan D, Pendergast DR. Quantitative effects of physical therapy on muscular and functional performance in subjects with osteoarthritis of the knees. Arch Phys Med Rehabil 1993;74: 840–7 (17).

Outcomes following drug trials for OA: Altman RD, Moskowitz R. Intraarticular sodium hyaluronate (Hyalgan) in the treatment of patients with osteoarthritis of the knee: a randomized clinical trial. J Rheumatol 1998;25:2203–12 (18).

Altman RD, Rosen JE, Bloch DA, Hatoum HT, Korner P. A double-blind, randomized, saline-controlled study of the efficacy and safety of EUFLEXXA for treatment of painful osteoarthritis of the knee, with an open-label safety extension (the FLEXX trial). Semin Arthritis Rheum 2009; 39:1–9 (19).

Predictive studies (risk/prevention) in hip and knee OA: Kauppila AM, Kyllonen E, Mikkonen P, Ohtonen P, Laine V, Siira P, et al. Disability in end-stage knee osteoarthritis. Disabil Rehabil 2009;31:370–80 (20).

Pua YH, Clark RA, Bryant AL. Physical function in hip osteoarthritis: relationship to isometric knee extensor steadiness. Arch Phys Med Rehabil 2010;91:1110–6 (21).

Thomas SG, Pagura SM, Kennedy D. Physical activity and its relationship to physical performance in patients with end stage knee osteoarthritis. J Orthop Sports Phys Ther 2003;33:745–54 (22).

Used in a number of different performance batteries: Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. Scand J Med Sci Sports 2001;11:280–6 (6).

Wright AA, Cook CE, Baxter GD, Garcia J, Abbott JH. Relationship between the Western Ontario and McMaster Universities Osteoarthritis Index Physical Function Subscale and physical performance measures in patients with hip osteoarthritis. Arch Phys Med Rehabil 2010;91: 1558–64 (12).

Cecchi F, Molino-Lova R, Di Iorio A, Conti AA, Mannoni A, Lauretani F, et al. Measures of physical performance capture the excess disability associated with hip pain or knee pain in older persons. J Gerontol A Biol Sci Med Sci 2009;64:1316–24 (23).

McCarthy CJ, Oldham JA. The reliability, validity and responsiveness of an aggregated locomotor function (ALF) score in patients with osteoarthritis of the knee. Rheumatology (Oxford) 2004;43:514–7 (24).

Rejeski WJ, Ettinger WH Jr, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995;3:157–67 (25).

Shields RK, Enloe LJ, Evans RE, Smith KB, Steckel SD. Reliability, validity, and responsiveness of functional tests in patients with total joint replacement. Phys Ther 1995; 75:169–79 (26).

Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 2006;86:1489–96 (27).

Steultjens MP, Roorda LD, Dekker J, Bijlsma JW. Responsiveness of observational and self-report methods for assessing disability in mobility in patients with osteoarthritis. Arthritis Rheum 2001;45:56–61 (28).

Juhakoski R, Tenhonen S, Anttonen T, Kauppinen T, Arokoski JP. Factors affecting self-reported pain and physical function in patients with hip osteoarthritis. Arch Phys Med Rehabil 2008;89:1066–73 (29) (Table 1).

Practical Application

How to obtain. No formal instructions required.

Method of administration. Performance based (assessed directly as test is performed). Equipment required: marked walkway, stopwatch, and tape measure. Assistive devices permissible (needs to be recorded and kept constant for comparisons).

Scoring. Time (seconds) measured with a stopwatch. Lower values represent better performance. Time can be converted to a walking speed by dividing the distance covered by the time taken. Usually expressed as meters/second. Higher values represent better performance.

Score interpretation. Mean ± SD normative speed reference values for healthy adults (30): woman age 50–59 years, 1.40 ± 0.15 meters/second (height normalized to 0.86); man age 50–59 years, 1.39 ± 0.23 meters/second (height normalized to 0.78); woman age 60–69 years, 1.28 ± 0.18 meters/second (height normalized to 0.81); man age 60–69 years, 1.36 ± 0.21 meters/second (height normalized to 0.78); woman age 70–79 years, 1.27 ± 0.21 meters/second (height normalized to 0.81); and man age 70–79 years, 1.33 ± 0.20 meters/second (height normalized to 0.76).

Gait speeds <1 meter/second identify a high risk of poor health-related outcomes in well-functioning older people (31). Older adults (ages ≥ 70 years) with slow gait speeds (≤ 0.7 meters/second) had a 1.5-fold increased risk of falls compared with those with normal speed (32).

Respondent burden. Minimal; <5 minutes.

Administrative burden. Minimal; <5 minutes. Time (seconds) is recorded on immediate completion of test. No training is required. Only 1 tester is required.

Translations/adaptations. Easy translated/adapted into any language.

Table 1. Physical performance battery membership for each physical performance measure*

		Physical performance battery									
Physical performance measure	ICF core activity	Shields et al, 1995 (26)	Rejeski et al, 1995 (25)	McCarthy and Oldham, 2004 (24)	Steultjens et al, 2001 (28)	Lin et al, 2001 (6)	Stratford and Kennedy, 2006 (84)	Juhakoski et al, 2008 (29)	Cecchi et al, 2009 (23)	Wright et al, 2010 (12)	French et al, in press (85)
SPWT (time)	d450: walking short distance d455: moving around	X		X	X	X	X	X	X	X	
SCT (time)	d410: changing body position d455: moving around d4551: climbing	X	X	X		X	X				
6MWT (distance)	d450: walking long distance d455: moving around		X		X		X	X	X		X
CST (count)	d410: changing body position			X	X	X			X	X	X
TUG (time)	d450: walking short distance d455: moving around d410 changing body position						X	X		X	X
LCT (time)	d430: lift and carry objects d450: walking short distance		X		X						
Sock Test (0–3 grade)	d540: dressing							X		X	
Car Task (time)	d410: changing body position		X								
Transfer to/from lying down (time)	d410: changing body position	X			X						
Step Test (count)	d410: changing body position									X	
Standing balance (time)	d415: maintaining body position								X		

* ICF = International Classification of Functioning, Disability and Health; SPWT = Self-Paced Walk Test; SCT = Stair Climb Test; 6MWT = Six-Minute Walk Test; CST = Chair Stand Test; TUG = Timed Up & Go; LCT = Lift and Carry Test.

Psychometric Information

Acceptability. Easy to use. Floor and ceiling effects have not been reported.

Reliability. *Evidence for internal consistency.* N/A.

Evidence for stability (test–retest). 50-feet SPWT: intrasession intraclass correlation coefficient ($ICC_{1,1}$) 0.97 (95% confidence interval [95% CI] 0.95, 0.98) in 82 people with end-stage hip and knee OA (8). 40-meter SPWT: long interval (median 178 days) $ICC_{2,1}$ 0.91 (95% CI 0.81, 0.97) in 21 people with end-stage hip and knee OA (11). 8-meter SPWT: intrasession $ICC_{2,1}$ 0.93 (95% CI 0.88, 0.99) in 41 people with knee OA. Significant differences were found between the 5 trials, and reliability improved when the first trial was discarded: $ICC_{2,1}$ 0.96 (95% CI 0.93, 0.98) with no significant differences found between trials (9). 8-meter SPWT: 1-week interval $ICC_{2,1}$ 0.88 (95% CI 0.78, 0.93), which increased with subsequent testing a week later ($ICC_{2,1}$ 0.94; 95% CI 0.89, 0.97) (9).

Evidence for interrater reliability. 50-feet SPWT: baseline $ICC_{1,1}$ 0.94 (95% CI 0.90, 0.96) in 82 people with end-stage hip and knee OA. Reliability improved on subsequent testing occasions: $ICC_{1,1}$ 0.96 (95% CI 0.93, 0.98) (8). 40-meter SPWT: baseline $ICC_{2,1}$ 0.95 (95% CI 0.90, 0.98) in 29 people with hip OA (mean ± SD age 66.5 ± 9.4 years) when tested within a 7-day period (13).

Measurement error: minimum detectable change at 90% confidence (MDC_{90}) and/or standard error of measurement (SEM). 40-meter SPWT: SEM 0.1 meter/second in 29 adults (mean ± SD age 66.5 ± 9.4 years) with hip OA (13). 50-feet SPWT: SEM 1.32 seconds (0.09 meter/second) and

MDC_{90} 3.08 seconds (0.2 meter/second) in 82 older people (mean ± SD age 70.3 ± 9.8 years) with end-stage hip and knee OA awaiting arthroplasty. This represents an 8.5% difference from trial to trial (8). 40-meter SPWT: SEM 1.73 seconds (95% CI 1.39–2.29; 0.14 meter/second) and MDC_{90} 4.04 seconds (0.33 meter/second) in 21 people with end-stage hip and knee OA when tested over a long interval (median 178 days) (11).

Validity. *Evidence of content validity.* N/A.

Evidence of construct validity: function. The 13-meter SPWT had a positive correlation with the Index of Severity for Knee (r = 0.66) in people with knee OA (7) and with the Lower Extremity Functional Scale (r = 0.59; 95% CI 0.44, 0.71) in people with knee and hip OA (33).

Evidence of construct validity: ROM. A positive correlation was found between walking speed and flexion range of the hip and knee among 702 community-dwelling older people (r = 0.40 and 0.35, respectively) (34).

Evidence of criterion validity. As baseline scores on the 8-feet SPWT decreased from 0.80 meter/second to 0.43 meter/second in community-dwelling older people, a greater percentage of people had disability in activities of daily living 4 years later (35). Adults ages ≥70 years with slow gait speeds (≤70 cm/second) had a 1.5-fold increased risk for falls (32).

Ability to detect change. *Evidence of responsiveness: standardized response means (SRMs) and effect sizes.* The 40-meter SPWT was responsive to detecting initial deterioration (n = 115; SRM −0.89 [95% CI −1.42, 0.68]) and then subsequent improvement (n = 89; SRM 0.79 [95% CI 0.66–1.45]) in the early postoperative period following hip or knee arthroplasty (11). Walking speed was not sensitive to change based on measures of disease activity in people with RA (36,37).

Interpretability: minimum clinically important differences (MCIDs). A small MCID of 0.05 meter/second and a substantial MCID of 0.10 meter/second were estimated for walking speed >10 meters in 492 community-dwelling elderly patients with mobility dysfunction, using both distribution- and anchor-based methods. Based on responsiveness indices, per-group sample size estimations when using gait velocity as an outcome measure were calculated as 142–161 subjects for small meaningful change and 37–42 subjects for substantial change (38).

In a sample of 65 patients with hip OA undergoing physiotherapy treatment, a comparison of 3 different anchor-based methods used to calculate MCIDs found that an increase greater than or equal to 0.2, 0.3, and 0.2 meter/second for the 40-meter SPWT was associated with a major improvement (defined as patient-reported change of greater than +5 on a −7 to +7 global rating of change scale) (13).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. A direct measure of walking speed over short distances, which is often limited in people with lower extremity OA. The test is quick and easy to perform and has minimal administrative or respondent burden. It appears to be relatively stable in people with lower extremity OA when tested over short durations (8) and appears to be a responsive outcome measure following rehabilitation and surgery (11). The SPWT has been used in 8 different performance batteries for people with lower extremity OA.

Caveats and cautions. Baseline practice effects have been found; therefore, a practice trial is necessary prior to baseline testing and should be considered for followup testing (8). External factors such as age, sex, and ethnicity can affect results on the SPWT (39,40). Depression and cognitive status are also associated with lower walking speeds (41). These external factors need to be considered when interpreting test results.

Clinical usability. Easy to administer, analyze, and interpret; readily available; requires little equipment; takes <5 minutes to perform; and can be conducted in most settings. It is recommended that a practice trial be provided and patients be monitored over several occasions to improve reliability (8,9).

The test–retest estimates of the SPWT met the requisite standards for making decisions at the individual patient level (11), whether tested by the same or different assessors (8). It is responsive to detecting deterioration and improvement in the early postoperative period (11).

The MDC_{90} of 0.2 meter/second found for the 50-feet SPWT provides information on the amount of change required to be confident that real change has occurred. In people with end-stage OA, this could represent up to 20% change (8).

Given that there is large variation in different methodologic approaches to define MCIDs, caution is needed when interpreting and using reported values to avoid misclassification of patient response to treatment (13).

Research usability. The SPWT is a reliable measure in lower extremity OA. It is recommended that a practice trial be provided and patients be monitored over several occasions to improve reliability (8,9).

Additional comparisons of methodologies used to calculate responsiveness and MCIDs are required in people with lower extremity OA (13).

STAIR CLIMB TEST (SCT)

Description

Purpose. The SCT assesses the ability to ascend and descend a flight of stairs, as well as lower extremity strength, power, and balance.

A number of test variations have been developed for different populations (osteoarthritis [OA], rheumatoid arthritis, elderly, cardiopulmonary, cerebrovascular accident, and children). Test variations in people with lower extremity OA include the number of steps, the task requirement (ascent only or ascent/decent combined), or whether the test is timed over a set number of steps or the step count is recorded for a set period of time: 9-step ascend/descend (11,25,42–46), 4-step ascend/descend (6), 3-step ascent cycle duration (47), 30-second test (12 steps) (48,49), 3-step measured on 0–6 scale of level of assistance (26), and 6-step fast and self-paced (21).

Content. A 9-step SCT was developed for end-stage hip and knee OA (11,42,45,46), which measures the time to ascend and descend a flight of 9 steps (step height 20 cm) in the usual manner at a safe and comfortable pace. Stepping pattern and use of aids are recorded. The use of a hand rail was not specified for this test.

A similar 5- or 9-step SCT was described for individuals with knee OA (25). The time required to ascend and descend an isolated set of 5 or 9 stairs (18 cm) using a single handrail is measured.

Domains covered. Mobility and climbing.

International Classification of Functioning, Disability and Health categories. d410–d429: changing and maintaining body position, d455: moving around, and d4551: climbing.

Number of items. 1.

Response options/scale. Time (seconds) or number of steps negotiated measured on a continuous ratio scale.

Recall period for items. Not applicable (N/A).

Endorsements. Cibulka MT, White DM, Woehrle J, Harris-Hayes M, Enseki K, Fagerson TL, et al. Hip pain and mobility deficits. Hip osteoarthritis: clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopaedic section of the American Physical Therapy Association. J Orthop Sports Phys Ther 2009;39:A1–25 (14).

Logerstedt DS, Snyder-Mackler L, Ritter RC, Axe MJ. Knee pain and mobility impairments: meniscal and articular cartilage lesions. J Orthop Sports Phys Ther 2010;40: A1–35 (50).

Examples of use. Outcome measure following hip/knee arthroplasty: Madsen OR, Brot C. Assessment of extensor and flexor strength in the individual gonarthrotic patient: interpretation of performance changes. Clin Rheumatol 1996;15:154–60 (49).

Farquhar S, Snyder-Mackler L. The Chitranjan Ranawat Award: the nonoperated knee predicts function 3 years after unilateral total knee arthroplasty. Clin Orthop Relat Res 2010;468:37–44 (51).

Floren M, Reichel H, Davis J, Laskin RS. The mini-incision mid-vastus approach for total knee arthroplasty. Oper Orthop Traumatol 2008;20:534–43 (52).

Zeni JA Jr, Snyder-Mackler L. Clinical outcomes after simultaneous bilateral total knee arthroplasty: comparison to unilateral total knee arthroplasty and healthy controls. J Arthroplasty 2010;25:541–6 (53).

Zeni JA Jr, Snyder-Mackler L. Early postoperative measures predict 1- and 2-year outcomes after unilateral total knee arthroplasty: importance of contralateral limb strength. Phys Ther 2010;90:43–54 (54).

Outcome measure following physical rehabilitation (exercise programs) in hip and knee OA: Galea MP, Levinger P, Lythgo N, Cimoli C, Weller R, Tully E, et al. A targeted home- and center-based exercise program for people after total hip replacement: a randomized clinical trial. Arch Phys Med Rehabil 2008;89:1442–7 (15).

Fisher NM, Gresham GE, Abrams M, Hicks J, Horrigan D, Pendergast DR. Quantitative effects of physical therapy on muscular and functional performance in subjects with osteoarthritis of the knees. Arch Phys Med Rehabil 1993;74: 840–7 (17).

Ettinger WH Jr, Burns R, Messier SP, Applegate W, Rejeski WJ, Morgan T, et al. A randomized trial comparing aerobic exercise and resistance exercise with a health education program in older adults with knee osteoarthritis: the Fitness Arthritis and Seniors Trial (FAST). JAMA 1997;277:25–31 (55).

Fransen M, Nairn L, Winstanley J, Lam P, Edmonds J. Physical activity for osteoarthritis management: a randomized controlled clinical trial evaluating hydrotherapy or Tai Chi classes. Arthritis Rheum 2007;57:407–14 (56).

Bennell KL, Hunt MA, Wrigley TV, Hunter DJ, McManus FJ, Hodges PW, et al. Hip strengthening reduces symptoms but not knee load in people with medial knee osteoarthritis and varus malalignment: a randomised controlled trial. Osteoarthritis Cartilage 2010;18:621–8 (57).

McKnight PE, Kasle S, Going S, Villanueva I, Cornett M, Farr J, et al. A comparison of strength training, self-management, and the combination for early osteoarthritis of the knee. Arthritis Care Res (Hoboken) 2010;62:45–53 (58).

Talbot LA, Gaines JM, Huynh TN, Metter EJ. A home-based pedometer-driven walking program to increase physical activity in older adults with osteoarthritis of the knee: a preliminary study. J Am Geriatr Soc 2003;51: 387–92 (59).

Outcomes following drug therapy trials: Clarke AK. A double-blind comparison of naproxen against indometacin in osteoarthrosis. Arzneimittelforschung 1975;25:302–4 (60).

Predictive studies (risk/prevention) in hip and knee OA: Pua YH, Clark RA, Bryant AL. Physical function in hip osteoarthritis: relationship to isometric knee extensor steadiness. Arch Phys Med Rehabil 2010;91:1110–6 (21).

Thomas SG, Pagura SM, Kennedy D. Physical activity and its relationship to physical performance in patients with end stage knee osteoarthritis. J Orthop Sports Phys Ther 2003;33:745–54 (22).

Singh JA, O'Byrne M, Harmsen S, Lewallen D. Predictors of moderate-severe functional limitation after primary total knee arthroplasty (TKA): 4701 TKAs at 2-years and 2935 TKAs at 5-years. Osteoarthritis Cartilage 2010;18: 515–21 (61).

Zeni JA Jr, Axe MJ, Snyder-Mackler L. Clinical predictors of elective total joint replacement in persons with end-stage knee osteoarthritis. BMC Musculoskelet Disord 2010;11:86 (62).

Zeni JA Jr, Snyder-Mackler L. Preoperative predictors of persistent impairments during stair ascent and descent after total knee arthroplasty. J Bone Joint Surg Am 2010; 92:1130–6 (63).

Cost-effectiveness: Sevick MA, Bradham DD, Muender M, Chen GJ, Enarson C, Dailey M, et al. Cost-effectiveness of aerobic and resistance exercise in seniors with knee osteoarthritis. Med Sci Sports Exerc 2000;32:1534–40 (64).

Used in a number of different performance batteries: Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. Scand J Med Sci Sports 2001;11:280–6 (6).

McCarthy CJ, Oldham JA. The reliability, validity and responsiveness of an aggregated locomotor function (ALF) score in patients with osteoarthritis of the knee. Rheumatology (Oxford) 2004;43:514–7 (24).

Rejeski WJ, Ettinger WH Jr, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995;3:157–67 (25).

Shields RK, Enloe LJ, Evans RE, Smith KB, Steckel SD. Reliability, validity, and responsiveness of functional tests in patients with total joint replacement. Phys Ther 1995; 75:169–79 (26).

Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 2006;86:1489–96 (27) (Table 1).

Practical Application

How to obtain. No formal instructions are required. Descriptions are available from the literature (25,42,46).

Method of administration. Performance based (assessed directly as test is performed). Equipment required: flight of 9–12 stairs and stopwatch to time in seconds to nearest tenth.

Scoring. Time (seconds) taken to complete the task, where smaller values represent better performance, and number of steps negotiated in set time, where larger values represent better performance.

Score interpretation. No formal normative values were found.

Respondent burden. Less than 5 minutes.

Administrative burden. Less than 5 minutes, including instructions. Time to score is upon completion of test where the time (seconds) or the number of steps negotiated is recorded. The step pattern (i.e., step-to-step, step-over-step) and use of gait aids and/or handrail can also be recorded. No training is required. Only 1 tester is required.

Translations/adaptations. Easy to translate/adapt into any language.

Psychometric Information

Acceptability. In a sample of 106 elderly people with symptomatic hip or knee OA, all were able to complete the SCT on 2 separate occasions (6).

Reliability. *Evidence for internal consistency.* N/A.

Evidence for stability (test–retest). 9-step SCT: long interval (median 178 days) intraclass correlation coefficient ($ICC_{2,1}$) 0.90 (95% confidence interval [95% CI] 0.79, 0.96) in 21 people with end-stage hip and knee OA (11). 4-step SCT: intrasession $ICC_{2,1}$ 0.94–0.96 (95% CI 0.75, 0.99) in 106 older adults with symptomatic hip and/or knee OA (6). 5- or 9-step SCT: r = 0.93 in 25 people with knee OA over a 2-week period (25). When tested over a 3-month period test–retest reliability decreased (r = 0.75), reflecting the possibility of true change over this timeframe (25).

Evidence for interrater reliability. Total interrater reliability between 3 clinicians: $ICC_{2,1}$ 0.94 (95% CI 0.55, 0.98) in 22 people 2–6 months following knee arthroplasty (65).

Measurement error: minimum detectable change at 90% confidence (MDC_{90}) and/or standard error of measurement (SEM). 9-step SCT: an SEM of 2.35 seconds (95% CI 1.89, 3.10) and an MDC_{90} of 5.5 seconds were found in a sample of 21 people with end-stage hip and knee OA awaiting arthroplasty (mean ± SD age 63.7 ± 10.7 years) (11). 4-step SCT: an SEM of 0.25–0.28 seconds was found in 106 elderly people with symptomatic hip/knee OA (6). 11-step SCT: an SEM of 1.14 seconds and an MDC_{90} of 2.6 seconds was found in a sample of 22 people with lower extremity OA following knee arthroplasty (65).

Validity. *Evidence of content validity.* N/A.

Evidence of construct validity: function. A positive correlation (r = 0.53) was found between the 4-step SCT and the Western Ontario and McMaster Universities Osteoarthritis Index physical functioning subscale in 106 elderly people with symptomatic hip/knee OA (6). A positive correlation (r = 0.44) was found between the 3-minute SCT (number of steps) and the Walking Impairment Questionnaire stair climbing subscale in 105 overweight patients with knee OA (66).

Evidence of construct validity: strength. Negative correlations were found between the 9-step SCT and quadriceps/hamstring strength in knee OA (43,67). Lower values (faster time) on the SCT indicated better performance, whereas higher values on the lower extremity strength test indicated better performance (r = −0.50 and −0.52, respectively).

Evidence of criterion validity. The SCT correlated well (r = 0.59–0.68) with other physical performance measures (figure-8 walk test, gait speed, chair rise test) when tested following knee arthroplasty. The SCT was best associated with gait speed and least associated with the chair rise test (65).

Ability to detect change. *Evidence of responsiveness: standardized response means (SRMs) and effect sizes (ES).* The 9-step SCT was responsive to detecting initial deterioration (n = 87; SRM −1.74 [95% CI −2.13, −1.45]) and then subsequent improvement (n = 73; SRM 1.98 [95% CI 1.68, 2.42]) in 150 patients during the early postoperative period following hip or knee arthroplasty (11). The 12-step SCT was the most responsive physical performance measure during short-term recovery following knee arthroplasty, with an ES of −0.71 (worsening) 1 month postoperatively and an ES of 0.84 (improvement) at 12-month followup (67).

Interpretability: minimum clinically important differences (MCIDs). No information on the MCID relevant to lower extremity OA could be found.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SCT is a direct measure of the ability to negotiate stairs, which is a common activity limitation and rehabilitation goal in people with lower extremity OA (68). The SCT appears to be stable in people with lower extremity OA and appears to be responsive to detect change (expected improvement and expected deterioration) following interventions such as physiotherapy and joint replacement surgery. The SCT correlates well with other physical performance measures and is best correlated with walking speed (6,65).

The SCT has been used in 6 different performance batteries for people with lower extremity OA (6,24–27,29).

Caveats and cautions. The SCT was dropped from 1 performance battery (27,69), as it was thought to be measuring more complex constructs than just physical performance similar to some of the self-reported measures (70).

Direct comparisons of the SCT across studies and during multiple assessments require the utilization of consistent specifications used during testing. These specifications include the step number, step height, and depth; use of a hand rail; and use of assistive devices.

Clinical usability. Easy to administer, analyze, and interpret; readily available; requires little equipment; takes <5 minutes to perform; and can be conducted in most settings provided access to stairs is available. There is a lack of information concerning the MCID for specific disease conditions, including OA.

Research usability. The SCT is relatively stable across time provided standardized environment, instructions, and encouragement are supplied. The SCT is found to have greater responsiveness than the patient-report questionnaires during the acute stages after knee replacement surgery and was shown to be the most responsive physical performance measure in the early recovery period (67). Further knowledge is required on the MCID in different disease conditions such as lower extremity OA.

SIX-MINUTE WALK TEST (6MWT)

Description

Purpose. The 6MWT assesses endurance and ability to walk over longer distances. The 6MWT was first described as a field test for physical fitness in 1963 (71) and then as a 12-minute walk test in people with chronic bronchitis (72). The 6MWT was found to perform as well as the 12-minute walk (73), and is now used to assess the submaximal level of functional performance at a similar level required for daily physical activities (74).

Used in many conditions, such as osteoarthritis (OA), cardiopulmonary disease, stroke, traumatic brain injury, patients who have undergone an amputation, Parkinson's disease, and Alzheimer's disease, as well as in elderly populations and children.

Content. Measures the distance an individual is able to walk in 6 minutes on a hard, flat, indoor surface. Standardized verbal encouragement (e.g., "5 minutes to go–keep going you are doing really well") can be provided at minute intervals and rest is allowed as required.

Domains covered. Walking long distances.

International Classification of Functioning, Disability and Health categories. d410–d429: changing and maintaining body position, d450: walking.

Number of items. 1.

Response options/scale. Distance (meters) measured on a continuous ratio scale.

Recall period for items. Not applicable (N/A).

Endorsements. The American College of Rheumatology (http://www.rheumatology.org/practice/clinical/clinician researchers/outcomes-instrumentation/6MWT.asp).

Logerstedt DS, Snyder-Mackler L, Ritter RC, Axe MJ. Knee pain and mobility impairments: meniscal and articular cartilage lesions. J Orthop Sports Phys Ther 2010;40: A1–35 (50).

Examples of use. Outcome measure following hip/knee arthroplasty: Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskelet Disord 2005;6:3 (11).

Parent E, Moffet H. Comparative responsiveness of locomotor tests and questionnaires used to follow early recovery after total knee arthroplasty. Arch Phys Med Rehabil 2002;83:70–80 (47).

Mizner RL, Petterson SC, Clements KE, Zeni JA Jr, Irrgang JJ, Snyder-Mackler L. Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments: a longitudinal analysis of outcomes. J Arthroplasty 2011;26: 728–37 (67).

Stratford PW, Kennedy DM, Maly MR, Macintyre NJ. Quantifying self-report measures' overestimation of mobility scores postarthroplasty. Phys Ther 2010;90:1288–96 (75).

Outcome measure following physical rehabilitation in hip and knee OA: Ettinger WH Jr, Burns R, Messier SP, Applegate W, Rejeski WJ, Morgan T, et al. A randomized trial comparing aerobic exercise and resistance exercise with a health education program in older adults with knee osteoarthritis: the Fitness Arthritis and Seniors Trial (FAST). JAMA 1997;277:25–31 (55).

Foley A, Halbert J, Hewitt T, Crotty M. Does hydrotherapy improve strength and physical function in patients with osteoarthritis: a randomised controlled trial comparing a gym based and a hydrotherapy based strengthening programme. Ann Rheum Dis 2003;62:1162–7 (76).

Deyle GD, Henderson NE, Matekel RL, Ryder MG, Garber MB, Allison SC. Effectiveness of manual physical therapy and exercise in osteoarthritis of the knee: a randomized, controlled trial. Ann Intern Med 2000;132: 173–81 (77).

Moffet H, Collet JP, Shapiro SH, Paradis G, Marquis F, Roy L. Effectiveness of intensive rehabilitation on functional ability and quality of life after first total knee arthroplasty: a single-blind randomized controlled trial. Arch Phys Med Rehabil 2004;85:546–56 (78).

Outcomes following drug therapy trials: Frestedt JL, Kuskowski MA, Zenk JL. A natural seaweed derived mineral supplement (Aquamin F) for knee osteoarthritis: a randomised, placebo controlled pilot study. Nutr J 2009; 8:7 (79).

Predictive studies (risk/prevention) in hip and knee OA: Juhakoski R, Tenhonen S, Anttonen T, Kauppinen T, Arokoski JP. Factors affecting self-reported pain and physical function in patients with hip osteoarthritis. Arch Phys Med Rehabil 2008;89:1066–73 (29).

Farquhar S, Snyder-Mackler L. The Chitranjan Ranawat Award: the nonoperated knee predicts function 3 years after unilateral total knee arthroplasty. Clin Orthop Relat Res 2010;468:37–44 (51).

Maly MR, Costigan PA, Olney SJ. Role of knee kinematics and kinetics on performance and disability in people with medial compartment knee osteoarthritis. Clin Biomech (Bristol, Avon) 2006;21:1051–9 (80).

Crosbie J, Naylor J, Harmer A, Russell T. Predictors of functional ambulation and patient perception following total knee replacement and short-term rehabilitation. Disabil Rehabil 2010;32:1088–98 (81).

Parent E, Moffet H. Preoperative predictors of locomotor ability two months after total knee arthroplasty for severe osteoarthritis. Arthritis Rheum 2003;49:36–50 (82).

Predictive studies for hospitalization and mortality: Lord SR, Menz HB. Physiologic, psychologic, and health predictors of 6-minute walk performance in older people. Arch Phys Med Rehabil 2002;83:907–11 (83).

Used in a number of different performance batteries for people with lower extremity OA: Cecchi F, Molino-Lova R, Di Iorio A, Conti AA, Mannoni A, Lauretani F, et al. Measures of physical performance capture the excess disability associated with hip pain or knee pain in older persons. J Gerontol A Biol Sci Med Sci 2009;64:1316–24 (23).

Rejeski WJ, Ettinger WH Jr, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995;3:157–67 (25).

Juhakoski R, Tenhonen S, Anttonen T, Kauppinen T, Arokoski JP. Factors affecting self-reported pain and physical function in patients with hip osteoarthritis. Arch Phys Med Rehabil 2008;89:1066–73 (29).

Stratford PW, Kennedy DM. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. J Clin Epidemiol 2006;59:160–7 (84).

French HP, Fitzpatrick M, FitzGerald O. Responsiveness of physical function outcomes following physiotherapy intervention for osteoarthritis of the knee: an outcome comparison study. Physiotherapy. In press (85) (Table 1).

Used as a measure on the Senior's Fitness Test (Fullerton Functional Test), developed as part of the LifeSpan Wellness Program: Rikli RE, Jones CJ. The development and validation of a functional fitness test for community-residing older adults. J Aging Phys Act 1999;7:129–61 (86).

Rikli RE, Jones CJ. Functional fitness normative scores for community-residing older adults, ages 60–94. J Aging Phys Act 1999;7:162–81 (87).

Practical Application

How to obtain. Descriptions are readily available online at: http://www.thoracic.org/statements/resources/pfet/sixminute.pdf and http://www.topendsports.com/testing/tests/walk-6min.htm.

Method of administration. Performance based (assessed directly as test is performed). Ideally conducted in an enclosed quiet hallway by a single administrator. A standardized procedure is important as performance can vary depending on the instructions provided, number of turns in the course, frequency and type of encouragement given, and number of trials performed.

Equipment required: 30 meters, premeasured flat walking area with interval markings every 3 meters, cones or brightly colored tape to mark boundaries of the course, watch or timer, and a chair (for resting if required).

Scoring. Distance (meters). Resting is allowed, but the time is not stopped. A greater distance represents better performance.

Score interpretation. In a sample of 109 (61 women) healthy white subjects ages 45–85 years, the average distances were: men, 682 meters (range 549–900) and women, 643 meters (range 479–816).

For patients with advanced lung disease/chronic obstructive pulmonary disease or heart failure, a 6MWT distance of $<$300 meters was associated with increased morbidity and mortality (88).

6MWT distances were found to be associated with age, sex, and height, and in women, body mass index (BMI). Regression equations to predict 6MWT in middle-aged and elderly adults were calculated (89): men, 6MWT (meters) = 867 − (5.71 age, years) + (1.03 height, cm) and women, 6MWT (meters) = 525 − (2.86 age, years) + (2.71 height, cm) − (6.22 BMI, kg/m^2).

Respondent burden. Minimal; $<$10 minutes. Could be physically demanding for very frail people or those with respiratory disorders.

Administrative burden. Less than 10 minutes, including instructions. Time to score is upon completion of the test where distance covered is calculated. It has been recommended that technicians who administer the 6MWT should be trained using a standard protocol, be supervised for several tests before administering them, and have cardiopulmonary resuscitation training (74).

Translations/adaptations. Easily translated/adapted into any language.

Psychometric Information

Acceptability. Possible ceiling effects for people with normal or high exercise capacities. Large baseline distances may limit the ability to detect performance improvements over time.

Reliability. *Evidence for internal consistency.* N/A.

Evidence for stability (test–retest). Intrasession stability intraclass correlation coefficient (ICC$_{2,1}$) 0.95–0.97 in 96 community-dwelling elderly people (ages 61–89 years) with independent functioning (40). Long interval (median 178 days) test–retest ICC$_{2,1}$ 0.94 (95% confidence interval [95% CI] 0.88, 0.98) in 21 people with end-stage hip and knee OA (11).

Measurement error: minimum detectable change at 90% confidence (MDC$_{90}$) and/or standard error of measurement (SEM). An SEM of 26.9 meters (95% CI 21.1, 34.8) and an MDC$_{90}$ of 61.3 meters were found in 21 people with end-stage hip and knee OA awaiting arthroplasty (mean ± SD age 63.7 ± 10.7 years) (11).

Validity. *Evidence of content validity.* N/A.

Evidence of construct validity: endurance. A positive correlation (r = 0.71) was found with maximum oxygen consumption following knee arthroplasty (90).

Evidence of construct validity: strength. Positive correlations (r = 0.44–0.47) were found with quadriceps/hamstring strength in knee OA (43,67) and (r = 0.58) with quadriceps strength 12 months post–knee arthroplasty (91).

Evidence of construct validity: function. A positive correlation was found with the Walking Impairment Questionnaire distance subscale (r = 0.52) and speed subscale (r = 0.51) in 105 overweight people with knee OA (66).

A positive correlation was found with the Short Form 36 physical function scale (r = 0.62) (91). The 6MWT was less well correlated with the Western Ontario and McMaster

Universities Osteoarthritis Index (WOMAC) physical function subscale (r = −0.27; a negative value was expected as greater values on the 6MWT indicated better performance, whereas lower values on the WOMAC indicated better function) in people with hip OA (29).

Evidence of criterion validity. Preoperative 6MWT scores, along with knee pain and knee flexion range, were significant predictors (adjusted $R^2 = 0.66$) of locomotor ability 2 months after knee arthroplasty (82).

Ability to detect change. *Evidence of responsiveness: standardized response means (SRMs) and effect sizes (ES).* The 6MWT was responsive to detecting initial deterioration (n = 82; SRM −1.74 [95% CI −1.97, −1.60]) and then subsequent improvement (n = 61; SRM 1.90 [95% CI 1.46, 2.39]) in the early postoperative period following hip or knee arthroplasty in 150 subjects (11). It was found to be the most responsive physical performance measure in 39 patients with knee OA following physiotherapy intervention (ES 0.43, SRM 0.54) (85).

Interpretability: minimum clinically important differences (MCIDs). A small MCID of 20 meters and a substantial MCID of 50 meters have been estimated for the 6MWT in a sample of 492 community-dwelling elderly people with mobility dysfunction when using both distribution- and anchor-based methods. Based on responsiveness indices, per-group sample size estimations for the 6MWT were calculated as 71–115 subjects for a small meaningful change and 13–20 subjects for a substantial meaningful change (38).

The smallest difference that was associated with a noticeable clinical difference in patients' perceptions of exercise performance (i.e., "a little bit better") was 54 meters (37,71) in a study of 112 patients (50% women) with stable severe chronic obstructive pulmonary disease (92).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The 6MWT measures submaximal functional performance, which is a common problem found in people with lower extremity OA (68). It appears to be sensitive to detect change following interventions such as physical therapy and joint replacement surgery. The measure appears to be appropriate for both early and end-stage OA as well as postarthroplasty.

The 6MWT has been used in a number of different performance batteries for people with lower extremity OA (23,25,29,84,85). It is also used in the Senior's Fitness Test (Fullerton Functional Test) developed as part of the LifeSpan Wellness Program (86,87).

Caveats and cautions. The 6MWT is a single measure that evaluates the global and integrated responses of the systems (cardiovascular, cardiopulmonary, and neuromuscular) involved during exercise. As such, a single distance score may not provide specific information on the function or contribution of each of the systems involved in the test. This may limit its use as an outcome measure for some populations, such as in systemic sclerosis (93).

A number of factors can cause variations in performance and therefore need to be documented. Encouragement has been shown to increase the distance walked (94). The number of trials performed can also vary the walk distance (95). It has been suggested that a practice test is not needed in most clinical settings but should be considered. If a practice test is performed, then at least 1-hour rest should be allowed before the second test. The greatest distance is then recorded (74).

Contraindications. Absolute: unstable coronary disease. Relative: resting heart rate >120 beats/minute, systolic blood pressure (BP) >180 mm Hg, diastolic BP >100 mm Hg; exertional angina without availability of antiangina medications; resting tachycardia >120 beats/minute; and syncope during exercise (especially in pulmonary hypertension) (74,88).

Clinical usability. Easy to administer, analyze, and interpret; readily available; requires little equipment; takes <10 minutes to perform; and can be conducted in most settings provided enough space is available.

There is a lack of information concerning the MCID for specific disease conditions, including OA. Current knowledge of MCIDs for the 6MWT in elderly people with mobility dysfunction suggests this value is smaller than the minimum detectable change (MDC) in people with lower extremity OA (38).

Research usability. The 6MWT is relatively stable across time provided standardized environment, instructions, and encouragement are supplied. Further knowledge is required on the MDC and MCID in different disease conditions such as lower extremity OA. Administrative and respondent burden does not limit research use.

CHAIR STAND TEST (CST)

Description

Purpose. The CST assesses the ability to rise from a chair and sit back down, as well as lower body strength and power.

Measured by either the time it takes to complete a specified number of chair stand repetitions (e.g., 10 repetitions [96], once [41], or 5 repetitions [35,97]) or the number of chair stand repetitions possible in a 30-second period (8,48,98).

The 10-repetition CST was originally developed for people with polymyositis and lower extremity weakness (96). A 30-second CST was later developed for community-dwelling older adults (ages 66–97 years) (98). It is also used in hip and knee osteoarthritis (OA), older adults, and children.

Content. For the 30-second CST, individuals are required to stand up from a standard chair (~43 cm) to a fully extended standing position as many times as possible with their arms folded across their chest. The number of completed repetitions achieved in 30 seconds is recorded (98).

Domains covered. Sitting and getting in/out of a seated position.

International Classification of Functioning, Disability and Health categories. d410: changing basic body position.

Number of items. 1.

Response options/scale. The total number of stand repetitions completed in 30 seconds or the time it takes to complete a specified number of repetitions measured on a continuous ratio scale.

Recall period for items. Not applicable (N/A).

Endorsements. None known.

Examples of use. Outcome measure following hip/knee arthroplasty: Boonstra MC, De Waal Malefijt MC, Verdonschot N. How to quantify knee function after total knee arthroplasty? Knee 2008;15:390–5 (99).

Catani F, Innocenti B, Belvedere C, Labey L, Ensini A, Leardini A. The Mark Coventry Award: articular contact estimation in TKA using in vivo kinematics and finite element analysis. Clin Orthop Relat Res 2010;468:19–28 (100).

Outcome measure following physical rehabilitation (exercise programs) in hip and knee OA: Talbot LA, Gaines JM, Huynh TN, Metter EJ. A home-based pedometer-driven walking program to increase physical activity in older adults with osteoarthritis of the knee: a preliminary study. J Am Geriatr Soc 2003;51:387–92 (59).

French HP, Fitzpatrick M, FitzGerald O. Responsiveness of physical function outcomes following physiotherapy intervention for osteoarthritis of the knee: an outcome comparison study. Physiotherapy. In press (85).

Arnold CM, Faulkner RA. The effect of aquatic exercise and education on lowering fall risk in older adults with hip osteoarthritis. J Aging Phys Act 2010;18:245–60 (101).

Piva SR, Gil AB, Almeida GJ, DiGioia AM 3rd, Levison TJ, Fitzgerald GK. A balance exercise program appears to improve function for patients with total knee arthroplasty: a randomized clinical trial. Phys Ther 2010;90:880–94 (102).

Wang C, Schmid CH, Hibberd PL, Kalish R, Roubenoff R, Rones R, et al. Tai Chi is effective in treating knee osteoarthritis: a randomized controlled trial. Arthritis Rheum 2009;61:1545–53 (103).

Outcomes following drug trials for OA: Fujita T, Fujii Y, Munezane H, Ohue M, Takagi Y. Analgesic effect of raloxifene on back and knee pain in postmenopausal women with osteoporosis and/or osteoarthritis. J Bone Miner Metab 2010;28:477–84 (104).

Predictive studies (risk/prevention) in hip and knee OA: Guralnik JM, Ferrucci L, Simonsick EM, Salive ME, Wallace RB. Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability. N Engl J Med 1995;332:556–61 (35).

The CST has been used in a number of different performance batteries: Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. Scand J Med Sci Sports 2001;11:280–6 (6).

Cecchi F, Molino-Lova R, Di Iorio A, Conti AA, Mannoni A, Lauretani F, et al. Measures of physical performance capture the excess disability associated with hip pain or knee pain in older persons. J Gerontol A Biol Sci Med Sci 2009;64:1316–24 (23).

McCarthy CJ, Oldham JA. The reliability, validity and responsiveness of an aggregated locomotor function (ALF) score in patients with osteoarthritis of the knee. Rheumatology (Oxford) 2004;43:514–7 (24).

Steultjens MP, Roorda LD, Dekker J, Bijlsma JW. Responsiveness of observational and self-report methods for assessing disability in mobility in patients with osteoarthritis. Arthritis Rheum 2001;45:56–61 (28).

French HP, Fitzpatrick M, FitzGerald O. Responsiveness of physical function outcomes following physiotherapy intervention for osteoarthritis of the knee: an outcome comparison study. Physiotherapy. In press (85).

Wright AA, Hegedus EJ, David Baxter G, Abbott JH. Measurement of function in hip osteoarthritis: developing a standardized approach for physical performance measures. Physiother Theory Pract 2011;27:253–62 (105) (Table 1).

Also used in the Senior's Fitness Test (Fullerton Functional Test) developed as part of the LifeSpan Wellness Program: Rikli RE, Jones CJ. The development and validation of a functional fitness test for community-residing older adults. J Aging Phys Act 1999;7:129–61 (86).

Rikli RE, Jones CJ. Functional fitness normative scores for community-residing older adults, ages 60–94. J Aging Phys Act 1999;7:162–81 (87).

Practical Application

How to obtain. Descriptions are readily available online at: http://www.topendsports.com/testing/tests/chair-stand.htm.

Method of administration. Performance based (assessed directly as test is performed). Equipment required: a straight back chair without arm rests (seat 43 cm high) and a stopwatch or timer (30 seconds).

Scoring. Number of repetitions (count), where higher values represent better performance, and time (seconds) taken to complete set number or repetitions, where smaller values (faster time) represent better performance.

Score interpretation. Normative scores for the 30-second CST in community-dwelling older people (87): age range 60–64 years, average count for women 12–17, average count for men 14–19; age range 65–69 years, average count for women 11–16, average count for men 12–18; age range 70–74 years, average count for women 10–15, average count for men 12–17; age range 75–79 years, average count for women 10–15, average count for men 11–17; age range 80–84 years, average count for women 9–14, average count for men 10–15; age range 85–89 years, average count for women 8–13, average count for men 8–14; age range 90–94 years, average count for women 4–11, average count for men 7–12.

The score interpretation calculator is available online at: http://www.exrx.net/Calculators/SeniorChairStand.html.

Respondent burden. Less than 3 minutes.

Administrative burden. Less than 3 minutes to administer, including instructions. Time to score is upon completion of test where the number of stands/time taken is recorded. The use of arms to assist in standing/sitting can also be recorded. No training is required. Only 1 tester is required.

Translations/adaptations. Easy to translate/adapt into any language.

Psychometric Information

Acceptability. Possible floor effects have been found for the repetition CST. In a sample of 106 older adults with symptomatic hip or knee OA, 24% were unable to complete the 5-repetition chair rise test due to pain (6).

Similarly, in a sample of 5,000 community-dwelling residents, 22% of people age >71 years could not complete the 5-repetition chair rise test (106).

Reliability. *Evidence for internal consistency.* N/A.

Evidence for stability (test–retest). 30-second CST: intrasession intraclass correlation coefficient ($ICC_{1,1}$) 0.95 (95% confidence interval [95% CI] 0.93, 0.97) in 82 older people with end-stage hip or knee OA (8). 5-repetition CST: intrasession $ICC_{2,1}$ 0.94 – 0.96 (95% CI 0.75, 0.99) in 106 elderly people with symptomatic hip and/or knee OA (6). 30-second CST: 2–5 days test–retest ICC 0.89 (95% CI 0.79, 0.936) in 76 community-dwelling older person (98).

Evidence of interrater reliability. 30-second CST: $ICC_{1,1}$ 0.93 (95% CI 0.87, 0.96) when tested at baseline, which improved on subsequent testing occasions ($ICC_{1,1}$ 0.98; 95% CI 0.96, 0.99) (8); and $ICC_{2,1}$ 0.81 (95% CI 0.63, 0.91) in 29 people with hip OA when tested within a 7-day period (13).

Measurement error: minimum detectable change at 90% confidence (MDC_{90}) and/or standard error of measurement (SEM). 30-second CST: an SEM of 1.3 repetitions was found in 29 people with hip OA (mean ± SD age 66.5 ± 9.4 years) (13); and an SEM of 0.7 repetitions (11% difference from trial to trial) and an MDC_{90} of 1.64 repetitions was found in 82 older people with end-stage hip and knee OA awaiting arthroplasty (mean ± SD age 70.3 ± 9.8 years) (8).

Validity. *Evidence of content validity.* N/A.

Evidence of construct validity: function. A positive correlation was found with walking speed (r = 0.66) in frail older adults (107) and a positive correlation was found with the Stair Climb Test (r = 0.59) when tested following knee arthroplasty (65).

Evidence of construct validity: strength. A moderate correlation was found with the weight-adjusted leg-press test of lower extremity strength in both community-dwelling older men (r = 0.78; 95% CI 0.63, 0.88) and women (r = 0.71; 95% CI 0.53, 0.84) (98).

Evidence of predictive validity. A 1-repetition CST >3.5 seconds was found to be a significant predictor of falls in the ambulatory frail older people, with an adjusted odds ratio of 3.4 (95% CI 1.2, 9.4) (108).

Ability to detect change. *Evidence of responsiveness: standardized response means (SRMs) and effect sizes (ES).* A small ES (0.36, SRM 0.39; mean change score 2.2 [95% CI 0.4, 4.1]) was found for the 5-repetition CST in 39 people with knee OA following outpatient exercised-based physiotherapy (85).

Interpretability: minimum clinically important differences (MCIDs). In a sample of 65 patients with hip OA undergoing physiotherapy treatment, a comparison of 3 different anchor-based methods used to calculate MCIDs found that an increase greater than or equal to 2.0, 2.6, and 2.1 repetitions on the 30-second CST was associated with a major improvement (defined as patient-reported change of greater than +5 on a −7 to +7 global rating of change scale) (13).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. A direct measure of the ability to rise from a chair, which is an activity that is commonly limited in people with lower extremity OA (4,42,68). The test is quick and easy to perform and has minimal administrative or respondent burden. The CST appears to be relatively stable in people with lower extremity OA when tested over short durations. The CST has been used in 6 different performance batteries for people with lower extremity OA.

Caveats and cautions. It is important to consider the CST version (i.e., 30 seconds or repetitions) being used, as psychometric information will be specific to the method. Baseline practice effects have been found; therefore, a practice trial is recommended prior to baseline testing and should be considered for followup testing (8). Floor effects have been noted for the repetition CST, and use of the 30-second CST may help overcome this (98). Other factors such as age, sex, and ethnicity can affect the results of the CST and need to be considered when interpreting test results (39,98).

Clinical usability. Easy to administer, analyze, and interpret. The CST requires little equipment, can be conducted in most settings, and takes <3 minutes to perform.

Available evidence of measurement error associated with the CST indicates that a change of at least 2 repetitions in a 30-second period is required to determine real change. In some age groups, this could require that up to 22% change is required to indicate real improvement or deterioration.

Given that there is large variation in different methodologic approaches to define MCIDs, caution is needed when interpreting and using reported values to avoid misclassification of patient response to treatment (13).

Research usability. Repeated intrasession testing may be limited, as repeated sit-to-stand activity may aggravate pain in people with OA (8). Limited information on ES limits the ability to estimate adequate sample sizes when using this measure as a primary outcome measure in clinical studies. Additional comparisons of methodologies used to calculate responsiveness and MCIDs are required in people with lower extremity OA (13).

TIMED UP & GO (TUG)

Description

Purpose. The TUG assesses basic mobility skill as well as strength, balance, and agility. Originally developed for frail elderly people as the “Get-Up and Go Test” in 1986 (109), it was adapted in 1991 to include the “time” component (110). The TUG is used in a range of populations from children to the elderly and for many conditions, including osteoarthritis (OA), joint arthroplasty, rheumatoid arthritis (RA), hip fractures, stroke, vertigo, and cerebral palsy.

Content. Time (seconds) taken to rise from sitting in an armchair, walk 3 meters, turn, walk back to the chair, then sit down using regular footwear and a walking aid if required.

Domains covered. Mobility and short distances (walking, turning, rising from a chair, sitting down into a chair).

International Classification of Functioning, Disability and Health categories. d410: changing basic body position, d450: walking, and d455: moving around.

Number of items. 1 (with 4 subcomponents: rising from sitting, walking, turning, and sitting back down).

Response options/scale. Timed (seconds) measured on a continuous ratio scale.

Recall period for items. Not applicable (N/A).

Endorsements. The American College of Rheumatology (http://www.rheumatology.org/practice/clinical/clinician researchers/outcomes-instrumentation/TUG.asp).

Cibulka MT, White DM, Woehrle J, Harris-Hayes M, Enseki K, Fagerson TL, et al. Hip pain and mobility deficits. Hip osteoarthritis: clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopaedic section of the American Physical Therapy Association. J Orthop Sports Phys Ther 2009;39:A1–25 (14).

Examples of use. Outcome measure following rehabilitation in hip and knee OA: Galea MP, Levinger P, Lythgo N, Cimoli C, Weller R, Tully E, et al. A targeted home- and center-based exercise program for people after total hip replacement: a randomized clinical trial. Arch Phys Med Rehabil 2008;89:1442–7 (15).

Fransen M, Nairn L, Winstanley J, Lam P, Edmonds J. Physical activity for osteoarthritis management: a randomized controlled clinical trial evaluating hydrotherapy or Tai Chi classes. Arthritis Rheum 2007;57:407–14 (56).

French HP, Fitzpatrick M, FitzGerald O. Responsiveness of physical function outcomes following physiotherapy intervention for osteoarthritis of the knee: an outcome comparison study. Physiotherapy. In press (85).

Hinman RS, Bennell KL, Crossley KM, McConnell J. Immediate effects of adhesive tape on pain and disability in individuals with knee osteoarthritis. Rheumatology (Oxford) 2003;42:865–9 (111).

Outcome measure following hip and knee arthroplasty: Zeni JA Jr, Axe MJ, Snyder-Mackler L. Clinical predictors of elective total joint replacement in persons with end-stage knee osteoarthritis. BMC Musculoskelet Disord 2010; 11:86 (62).

Mizner RL, Petterson SC, Clements KE, Zeni JA Jr, Irrgang JJ, Snyder-Mackler L. Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments: a longitudinal analysis of outcomes. J Arthroplasty 2011;26: 728–37 (67).

Boonstra MC, De Waal Malefijt MC, Verdonschot N. How to quantify knee function after total knee arthroplasty? Knee 2008;15:390–5 (99).

Kennedy DM, Hanna SE, Stratford PW, Wessel J, Gollish JD. Preoperative function and gender predict pattern of functional recovery after hip and knee arthroplasty. J Arthroplasty 2006;21:559–66 (112).

Outcomes of drug therapy trials in knee OA: Kraemer WJ, Ratamess NA, Anderson JM, Maresh CM, Tiberio DP, Joyce ME, et al. Effect of a cetylated fatty acid topical cream on functional mobility and quality of life of patients with osteoarthritis. J Rheumatol 2004;31:767–74 (113).

Predictive studies (risk/prevention) in hip and knee OA: Kennedy DM, Hanna SE, Stratford PW, Wessel J, Gollish JD. Preoperative function and gender predict pattern of functional recovery after hip and knee arthroplasty. J Arthroplasty 2006;21:559–66 (112).

Arnold CM, Faulkner RA. The history of falls and the association of the timed up and go test to falls and near-falls in older adults with hip osteoarthritis. BMC Geriatr 2007;7:17 (114).

Halket A, Stratford PW, Kennedy DM, Woodhouse LJ. Using hierarchical linear modeling to explore predictors of pain after total hip and knee arthroplasty as a consequence of osteoarthritis. J Arthroplasty 2010;25:254–62 (115).

Used in a number of different performance batteries: Wright AA, Cook CE, Baxter GD, Garcia J, Abbott JH. Relationship between the Western Ontario and McMaster Universities Osteoarthritis Index Physical Function Subscale and physical performance measures in patients with hip osteoarthritis. Arch Phys Med Rehabil 2010;91: 1558–64 (12).

Cecchi F, Molino-Lova R, Di Iorio A, Conti AA, Mannoni A, Lauretani F, et al. Measures of physical performance capture the excess disability associated with hip pain or knee pain in older persons. J Gerontol A Biol Sci Med Sci 2009;64:1316–24 (23).

Juhakoski R, Tenhonen S, Anttonen T, Kauppinen T, Arokoski JP. Factors affecting self-reported pain and physical function in patients with hip osteoarthritis. Arch Phys Med Rehabil 2008;89:1066–73 (29).

Stratford PW, Kennedy DM. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. J Clin Epidemiol 2006;59:160–7 (84) (Table 1).

Practical Application

How to obtain. Descriptions are readily available online at: http://www.unmc.edu/media/intmed/geriatrics/nebgec/pdf/frailelderlyjuly09/toolkits/timedupandgo_w_norms.pdf.

Method of administration. Performance based (assessed directly as test is performed). Simple test with minimal equipment: standard arm chair (seat height ~46 cm, arm height ~65 cm), 3-meter walkway with floor mark, and stopwatch or watch with time in seconds. It is recommended that 2 trials are performed and the best result is used.

Scoring. Time (seconds). Smaller values (faster time) represent better performance.

Score interpretation. Normative age group reference (116): age 60–69 years, time 8.1 seconds (95% confidence interval [95% CI] 7.1, 9.0); age 70–79 years, time 9.2 seconds (95% CI 8.2, 10.2); and age 80–99 years, time 11.3 seconds (95% CI 10.0, 12.7).

In frail elderly people, scores <10 seconds = normal; 10–19 seconds = good mobility, can go out alone, mobile without a gait aid; 20–29 seconds = problems, cannot go outside alone, requires a gait aid; and ≥30 seconds = with increased functional dependence (110). Older adults who take >14 seconds to complete the TUG have a high risk for falls (117).

Respondent burden. Minimal; <3 minutes.

Administrative burden. Minimal; <3 minutes. Score is immediate. No training is required. Only 1 tester is required.

Translations/adaptations. Easy to translate/adapt into any language. Adaptations: TUG extended (117), i.e., 1) TUG Cognitive: TUG test while counting backward from a randomly selected number between 20 and 100 and 2) TUG manual: TUG while carrying a full cup of water, and i-TUG (118), i.e., uses portable inertial sensors to automatically detect and separate the subcomponents of the TUG.

Psychometric Information

Acceptability. In 1,200 community-dwelling older people with varying cognitive and functional ability (including fallers), 6% refused to complete the test and none were unable to complete the test (119). All but 1 was able to complete the test in a study of 65 older people with RA (120).

A floor effect has been found in short-term hospitalized older people, with ~25% unable to complete the test (121). Ceiling effects after knee replacement have been found where improvements reach a plateau earlier than other physical performance measures (11).

Reliability. *Evidence for internal consistency.* N/A.

Evidence for stability (test–retest). Intrasession stability in 96 community-dwelling older people (ages 61–89 years) with independent functioning was high (intraclass correlation coefficient [$ICC_{2,1}$] 0.95–0.97) (40). Test–retest reliability in frail older people over longer periods (up to 132 days) was less reliable (ICC 0.74) (107). Test–test reliability in 21 people with end-stage hip and knee OA when tested over a long interval (median 178 days) was not sufficient for individual patient use ($ICC_{2,1}$ 0.75; 95% CI 0.51, 0.98) Testing over this time period was likely to have captured true change in this population (11).

Evidence for interrater reliability. $ICC_{2,1}$ 0.87 (95% CI 0.63, 0.91) in 29 people with hip OA (mean ± SD age 66.5 ± 9.4 years) when tested within a 7-day period (13). In the frail older people, interrater reliability was high within the same day (ICC 0.99) and on a consecutive visit (ICC 0.99) when measured by a physical therapist, physician, and patient attendant (110). In 22 people with RA, interrater reliability among 3 physical therapists was also high (r = 0.97) (122).

Measurement error: minimum detectable change at 90% confidence (MDC_{90}) and/or standard error of measurement (SEM). An SEM of 0.84 seconds was found in a sample of 29 people with hip OA (mean ± SD age 66.5 ± 9.4 years) (13). An SEM of 1.07 seconds (95% CI 0.86, 1.41) and an MDC_{90} of 2.49 seconds was found in a sample of 21 people with end-stage hip and knee OA awaiting arthroplasty (mean ± SD age 63.7 ± 10.7 years) (11). An SEM of 1 second was found in 22 people with RA (mean 60 years, range 18–80 years) when tested over a 2–7-day period (122).

Validity. *Evidence of content validity.* N/A.

Evidence of construct validity: balance. Negative correlations were found with the Berg Balance Scale in frail older people (r = −0.81) (110) and with the Tinetti Performance-Oriented Mobility Assessment balance in community dwelling older people (r = −0.55) (119).

Evidence of construct validity: gait speed. A negative correlation with gait speed has been found in frail older people (r = −0.61) (110).

Evidence of construct validity: strength. A negative correlation with quadriceps strength (r = −0.49) and hamstring strength (r = −0.51) has been found in people with knee OA (123).

Evidence of construct validity: function. A negative correlation with the Bartel Index of Activity has been found in older people (r = −0.78) (110). Not well correlated with the Western Ontario and McMaster Universities Osteoarthritis Index physical function subscale (r = 0.282) in patients with hip OA, as they are measuring different constructs (29,84). Selective in discriminating between 28 knee patients following total knee arthroplasty for OA and 31 healthy controls (99).

Evidence of criterion validity. A TUG >10 seconds was predictive of near falls in older people with hip OA (odds ratio [OR] 3.1; 95% CI 1.0, 9.9) (114). A preoperative TUG ≥15.3 seconds was sensitive (83.3%) and specific (61.1%) to predict a deep vein thrombosis in a sample of 38 patients with hip OA following a total hip arthroplasty (OR 7.0; 95% CI 1.6, 30.8) (124).

Ability to detect change. *Evidence of responsiveness: standardized response means (SRMs) and effect sizes (ES).* Responsive in detecting initial deterioration (n= 116; SRM −1.08 [95% CI −1.38, 0.92]) and then subsequent improvement (n = 89; SRM 1.04 [95% CI 0.84, 1.61]) in the early postoperative period following hip or knee arthroplasty in 150 subjects (11).

Improvement in TUG time ($P = 0.01$) was found following a quadriceps and hamstrings strengthening program in a 36 people with RA (125).

A small ES (0.33, SRM 0.35) was found in 39 patients with knee OA following physiotherapy intervention. Median change score for this knee OA population was 1 second (95% CI 0.1, 1.9) (85).

Sensitivity of the TUG to detect change where change has occurred has been questioned in the less severe OA patients (126).

Interpretability: minimum clinically important differences (MCIDs). In a sample of 65 patients with hip OA undergoing physiotherapy treatment, a comparison of 3 different anchor-based methods used to calculate MCIDs found that a reduction greater than or equal to 0.8, 1.4, and 1.2 seconds on the TUG was associated with a major improvement (defined as patient-reported change of greater than +5 on a −7 to +7 global rating of change scale) (13).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Easy to administer and can be used in most environmental contexts. Requires minimal equipment and interpretation. The TUG assesses common problems found in people with lower extremity OA and can be used for a variety of other populations. Appears to be a responsive outcome measure following rehabilitation and surgery.

Has been used a in a number of different performance batteries (12,23,29,84).

Caveats and cautions. The stability of the TUG over longer intervals would not meet the standards for individual patient use (11). The TUG is limited for cognitively impaired frail older people, as up to 35.5% of this subpopulation is unable to physically perform the test (127).

As the TUG incorporates 4 different subcomponents that represent different functioning constructs, the total score (time in seconds) limits interpretation about the proportional contribution of these subcomponents on activity limitation.

Clinical usability. Easy to administer, analyze, and interpret; readily available; requires little equipment; takes <5 minutes to perform; and can be conducted in most settings. Research provided about MDC and SRM provides information on outcomes following intervention and true change over time (11).

Given that there is large variation in different methodologic approaches to define MCIDs, caution is needed when interpreting and using reported values to avoid misclassification of patient response to treatment (13).

Floor and ceiling effects may limit the use of the TUG directly following surgery such as joint replacement arthroplasty (24,25).

Research usability. Results may be affected by floor and ceiling effects in some subgroups. Interpretation may be limited by the multiple constructs contained in the measure. Administrative burden or respondent burden does not limit research use.

Additional comparisons of methodologies used to calculate responsiveness and MCIDs are required in people with lower extremity OA (13).

SOCK TEST

Description

Purpose. The Sock Test assesses the ability to put on a sock or footwear. Originally developed for people with musculoskeletal pain. Also used as an outcome measure in people with back pain (128) and hip osteoarthritis (OA) (12,29,105).

Content. Starting from sitting on a high bench, with feet off the floor, the person is instructed to lift up 1 leg at a time in the sagittal plane and simultaneously reach down toward the lifted foot with both hands, one on each side, grabbing the toes with the fingertips of both hands. The foot must not touch the bench and should be in the air at all times during the test. After testing each leg once, the patient is given a score on the most limited performance. Scores are given as ordinal values from 0 (can grab the toes with fingertips and perform the action with ease) to 3 (can hardly, if at all, reach as far as the malleoli).

Domains covered. Flexibility and putting on/taking off socks.

International Classification of Functioning, Disability and Health categories. d540: dressing.

Number of items. 1.

Response options/scale. The test is measured using a 4-point ordinal scale (0–3) reflecting the ability and ease of the performance.

Recall period for items. Not applicable (N/A).

Endorsements. None found.

Examples of use. Physical performance measure in a back pain performance battery: Strand LI, Moe-Nilssen R, Ljunggren AE. Back Performance Scale for the assessment of mobility-related activities in people with back pain. Phys Ther 2002;82:1213–23 (128).

Physical performance measure in OA performance batteries: Wright AA, Cook CE, Baxter GD, Garcia J, Abbott JH. Relationship between the Western Ontario and McMaster Universities Osteoarthritis Index Physical Function Subscale and physical performance measures in patients with hip osteoarthritis. Arch Phys Med Rehabil 2010;91:1558–64 (12).

Wright AA, Hegedus EJ, David Baxter G, Abbott JH. Measurement of function in hip osteoarthritis: developing a standardized approach for physical performance measures. Physiother Theory Pract 2011;27:253–62 (105).

Practical Application

How to obtain. Descriptions are available from the literature (129).

Method of administration. Performance based (assessed directly as test is performed). Equipment required: high bench.

Scoring. 4-point ordinal grading (0–3). Smaller values represent better performance: 0 = can grab toes with fingertips, perform the action with ease; 1 = can grab toes with fingertips, but performs action with effort; 2 = can reach beyond the malleoli, but cannot reach toes; and 3 = can hardly, if at all, reach as far as the malleoli.

Score interpretation. No formal normative values are available.

Respondent burden. Less than 5 minutes.

Administrative burden. Less than 5 minutes, including instructions. Time to score is upon completion of test where the test is graded on a 4-point ordinal scale.

Translations/adaptations. Easy to translate/adapt into any language.

Psychometric Information

Acceptability. In a study of 93 people with hip OA (mean age 66.4 years, range 41–85 years, and mean body mass index [BMI] 28.97 kg/m^2, range 20.37–48.72), all participants were able to participate in the Sock Test (12).

Reliability. *Evidence for internal consistency.* N/A.

Evidence for stability (test–retest). No evidence found.

Evidence of interrater reliability. 14-day interval: weighted $\kappa = 0.79$ (95% confidence interval 0.5, 1.0) in 21 people with musculoskeletal pain (129).

Measurement error: minimum detectable change and/or standard error of measurement. No evidence found.

Validity. *Evidence of content validity.* N/A.

Evidence of construct validity: function. A positive correlation was found with the Disability Rating Index ($r_s = 0.45$) in people with musculoskeletal pain (129). A weak correlation was found with the Western Ontario and McMaster Universities Osteoarthritis Index physical functioning subscale (r = 0.243) (29).

Evidence of criterion validity. An increased likelihood of patient-perceived functional difficulty (measured using yes/no answers to a set of 3 questions) at 1-year followup was found with higher pretest Sock Test scores (129). Using a score of 0 as a reference, a score of 2 on the Sock Test increased the likelihood of perceived functional difficulties after 1 year by 6 times, and a score of 3 increased this by 12 times (129).

The sensitivity and specificity of the Sock Test to patient-reported (yes/no) activity limitation in 237 adults with musculoskeletal pain (129) were reported as: Sock Test = 1, sensitivity 0.77, specificity 0.91; Sock Test = 2, sensitivity 0.99, specificity 0.31; Sock Test = 3, sensitivity 1.00, specificity 0.25.

Ability to detect change. *Evidence of responsiveness: standardized response means and effect sizes.* Changes in Sock Test scores between baseline testing and 1-year followup testing correlated best with changes in the dressing items on the Disability Rating Index, but the correlation was low (dressing items = 0.36, overall = 0.35) (129).

Interpretability: minimum clinically important differences (MCIDs). No relevant evidence for MCID was found.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. A direct measure of putting on and taking off socks and footwear, which is a commonly reported problem in people with lower extremity OA (1). The test is quick and easy to perform and has minimal administrative or respondent burden.

The Sock Test has been used in 2 different performance batteries for people with lower extremity OA (12,29) (Table 1).

Caveats and cautions. Stability of this measure has not been determined and evidence to the responsiveness to change is limited. There is an increased likelihood ($P <$ 0.05) of scoring 1 or higher on the Sock Test with increases in age and BMI. Patients between ages 51 and 65 years were almost 3 times more likely to score >0 on the Sock Test than patients between ages 21 and 35 years. Patients with BMI values >27.1 kg/m^2 were almost 10 times more likely to score >0 on the Sock Test than patients with BMI <22.1 kg/m^2 (129).

Clinical usability. Easy to administer, analyze, and interpret; readily available; requires little equipment; takes <3 minutes to perform; and can be conducted in most settings provided a sufficiently high enough bench is available.

There is a lack of information concerning the stability, measurement error, and MCID for specific disease conditions, including lower extremity OA, limiting the interpretability in the clinical setting.

Research usability. Further evaluation of the Sock Test is required to determine the stability, measurement error, and MCID for disease-specific populations, including people with OA.

LIFT AND CARRY TEST (LCT)

Description

Purpose. The LCT assesses how quickly and how easily a person can lift up and carry an object over a short distance. It was developed in 1995 for people with knee osteoarthritis (OA) (25). Variations on the LCT have been used in other study populations such as lower back pain (128).

Content. The test requires the person to walk ~2.7 meters to a set of shelves, pick up a 4.5 kg weight (reported in error as 22 kg in the original article; correct weight of 4.5 kg [10 lbs] was confirmed with the corresponding author) from the lower shelf (approximately knee height), turn and carry the weight for ~4.35 meters around a cone, return to the shelves, and place the weight on a high shelf (approximately shoulder height) as quickly as possible.

Domains covered. Lifting and carrying objects and walking short distances.

International Classification of Functioning, Disability and Health categories. d430: lift and carry objects and d450: walking short distances.

Number of items. 1.

Response options/scale. Time (seconds). Measured on a continuous ratio scale. The task is also rated on a self-perceived difficulty 0–10 ordinal scale.

Recall period for items. Not applicable (N/A).

Endorsements. None found.

Examples of use. Outcome measure following physical rehabilitation (exercise programs) in knee OA: Ettinger WH Jr, Burns R, Messier SP, Applegate W, Rejeski WJ, Morgan T, et al. A randomized trial comparing aerobic exercise and resistance exercise with a health education program in older adults with knee osteoarthritis: the Fitness Arthritis and Seniors Trial (FAST). JAMA 1997;277: 25–31 (55).

Predictive studies (risk/prevention) in hip and knee OA: Bieleman HJ, Reneman MF, van Ittersum MW, van der Schans CP, Groothoff JW, Oosterveld FG. Self-reported functional status as predictor of observed functional capacity in subjects with early osteoarthritis of the hip and knee: a diagnostic study in the CHECK cohort. J Occup Rehabil 2009;19:345–53 (130).

Physical performance measure in a performance battery for lower extremity OA: Rejeski WJ, Ettinger WH Jr, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995;3:157–67 (25) (Table 1).

Practical Application

How to obtain. Full instructions, including script and equipment, are available from the original publication (25).

Method of administration. Performance based (assessed directly as test is performed). Equipment required: shelves (low set to knee height and high set at shoulder height), weighted object, stopwatch, marked floor, and cone.

Scoring. Time (seconds). Timing stops when the weight first touches the high shelf. Faster (lower) times represent better performance. The person then rates the self-perceived demand of the task on a 0–10-point difficulty scale, where 0 = easy and 10 = very difficult.

Score interpretation. No information found.

Respondent burden. Less than 5 minutes.

Administrative burden. Less than 5 minutes, including instructions. Time to score is upon completion of test where the time (seconds) is recorded and the self-reported demand of the task is recorded on a 0–10-point difficulty scale.

Translations/adaptations. None found.

Psychometric Information

Method of development. Single item that evaluates the ability to lift and carry a weight over short distances as well as rate the self-perceived demand of the task.

Acceptability. It is unknown whether there are any floor or ceiling effects.

Reliability. *Evidence for internal consistency.* N/A.

Evidence for stability (test–retest). Stability of over a 14-day period: r = 0.92 in 25 people with knee OA (25). Stability over a 3-month period: r = 0.77 in 72 people with hip and knee OA. Stability may have been compromised by real change following a health education intervention that occurred in the interim period (25).

Measurement error: minimum detectable change (MDC) and/or standard error of measurement (SEM). An SEM of 0.49–0.50 second in a sample of 25 people with knee OA and 0.27–0.30 second in a sample of 78 people with knee OA can be calculated from the data reported by Rejeski et al (25).

Validity. *Evidence of content validity.* N/A.

Evidence of construct validity. Significant correlation with the Fitness and Arthritis Trial Functional Activities Inventory ambulation and climbing subscale (r = 0.34) (25).

Evidence of criterion validity. As hypothesized, lower values (faster times) on the LCT were correlated (negative correlations were expected as lower values [faster time] on the LCT indicate better performance, whereas higher values on the treadmill, peak oxygen consumption [$\mathrm{Vo}_{2\max}$], and knee strength tests indicated better performance) concurrently with higher values on the treadmill time (r = −0.40), $\mathrm{Vo}_{2\max}$ (r = −0.38), and knee strength (r = −0.58) in 104–437 people with knee OA (25).

Ability to detect change. *Evidence of responsiveness: standardized response means and effect sizes (ES).* Responsive to the effects of an aerobic exercise intervention (mean ± standard error of the mean 9.0 ± 0.2 seconds versus 10.0 ± 0.1 seconds; $P < 0.001$) and a resistance exercise intervention (mean ± standard error of the mean 9.3 ± 0.1 seconds versus 10.0 ± 0.1 seconds; $P = 0.003$) (55). (Note: only ± standard error of the mean was used, not SD, and ES cannot be calculated).

Interpretability: minimum clinically important differences (MCIDs). No relevant evidence was found.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. A direct measure of lifting and carrying activities (such as carrying groceries or washing), which are reported to be limited in people with lower extremity OA (68). Quick and easy to perform, minimal administrative or respondent burden, but requires a contextual setup. Appears to be stable over shorter durations (2 weeks); however, may be influenced by external factors and true change over longer periods (3 months). The LCT has been used in 1 performance battery for people with lower extremity OA.

Caveats and cautions. Small systematic improvements across time have been noted, which may reflect motivational or learning effects (25).

Clinical usability. Easy to administer, analyze, and interpret; takes <5 minutes to perform; and can be conducted in most settings provided access to the appropriate equipment is available. There is a lack of information concerning the MDC, SEM, and MCID for specific disease conditions, including OA.

Research usability. More research is required to determine the responsiveness to change, MDCs, and MCIDs in people with OA and other populations.

CAR TASK

Description

Purpose. The Car Task assesses how quickly and easily a person can get in and out of the car. It was originally developed for people with knee osteoarthritis (OA) (25).

Content. Starting ~27 cm away from the car door with the hip aligned with the edge of the door, participants are instructed to open the door, sit down in the car, close the door, reopen the door, and then step out to resume a fully erect standing position.

Domains covered. Getting in/out of a car.

International Classification of Functioning, Disability and Health categories. d410: changing basic body position and d475: driving.

Number of items. 1.

Response options/scale. Time (seconds) measured on a continuous ratio scale. The task is also rated on a self-perceived difficulty 0–10 ordinal scale.

Recall period for items. Not applicable (N/A).

Endorsements. None found.

Examples of use. Outcome measure following physical rehabilitation (exercise programs) in knee OA: Ettinger WH Jr, Burns R, Messier SP, Applegate W, Rejeski WJ, Morgan T, et al. A randomized trial comparing aerobic exercise and resistance exercise with a health education program in older adults with knee osteoarthritis: the Fitness Arthritis and Seniors Trial (FAST). JAMA 1997;277: 25–31 (55).

Predictive studies (risk/prevention) in hip and knee OA: Bieleman HJ, Reneman MF, van Ittersum MW, van der Schans CP, Groothoff JW, Oosterveld FG. Self-reported functional status as predictor of observed functional capacity in subjects with early osteoarthritis of the hip and knee: a diagnostic study in the CHECK cohort. J Occup Rehabil 2009;19:345–53 (130).

Physical performance measure in a performance battery for lower extremity OA: Rejeski WJ, Ettinger WH Jr, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995;3:157–67 (25) (Table 1).

Practical Application

How to obtain. Full instructions, including script and equipment, are available from the original publication (25).

Method of administration. Performance based (assessed directly as test is performed). Equipment required: car, marker, and stopwatch.

Scoring. Time (seconds) starts at the command "Go" and stops when the person is standing fully erect. Smaller values (faster time) represent better performance. The person then rates the self-perceived demand of the task on a 0–10-point difficulty scale, where 0 = easy and 10 = very difficult.

Score interpretation. No information found.

Respondent burden. Less than 10 minutes.

Administrative burden. Less than 10 minutes, including instructions. Time to score is upon completion of the test where the time (seconds) is recorded and the self-reported demand of the task is recorded on a 0–10-point difficulty scale.

Translations/adaptations. None found.

Psychometric Information

Acceptability. It is unknown whether there are any floor or ceiling effects.

Reliability. *Evidence for internal consistency.* N/A.

Evidence for stability (test–retest). 14-day test–retest: r = 0.88 in 25 people with knee OA (25). 3-month test–retest: r = 0.86 (25).

Measurement error: minimum detectable change (MDC) and/or standard error of measurement (SEM). An SEM of 0.88–0.97 second in a sample of 25 people with knee OA can be calculated by data reported by Rejeski et al (25).

Validity. *Evidence of content validity.* N/A.

Evidence of construct validity. Significant correlation with the Fitness and Arthritis Trial Functional Activities Inventory: ambulation and climbing subscale (r = 0.38) and complex activities subscale (r = 0.35) (25).

Evidence of criterion validity. When tested concurrently, lower values (faster times) on the Car Task were significantly correlated (negative correlations were expected as lower values [faster time] on the Car Task indicate better performance, whereas higher values on the treadmill, peak oxygen consumption [$\text{V}\text{O}_{2\text{peak}}$], and knee strength tests indicated better performance) with higher values on treadmill time (r = −0.45), $\text{V}\text{O}_{2\text{peak}}$ (r = −0.40), and knee strength (r = −0.46) when tested in 104–209 people with knee OA (25).

Ability to detect change. *Evidence of responsiveness: standardized response mean and effect sizes (ES).* Responsive to the effects of an aerobic exercise intervention (mean ± standard error of the mean 8.7 ± 0.4 seconds versus 10.6 ± 0.3 seconds; $P < 0.001$) and a resistance exercise intervention (mean ± standard error of the mean 9.0 ± 0.3 seconds versus 10.6 ± 0.3 seconds; $P = 0.003$) (Note: only ± standard error of the mean was used, not SD, and ES cannot be calculated).

Interpretability: minimum clinically important differences (MCIDs). No relevant evidence was found.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Direct measure of the ability to get into and out of a car, which is reported to be limited in people with lower extremity OA (25). The test is quick to perform and has minimal administrative or respondent burden, but requires a contextual setup (i.e., access to a car). Initial testing appears to be stable over shorter durations (2 weeks); however, it may be influenced by external factors, such as learning effects over longer periods (3 months). The Car Task has been used in 1 performance battery for people with lower extremity OA (25) (Table 1).

Caveats and cautions. Small systematic improvements across time have been noted, which may reflect motivational or learning effects (25).

Clinical usability. Easy to administer, analyze, and interpret and takes <10 minutes to perform, but requires outdoor access to a car. There is a lack of information concerning the measurement error, responsiveness to change, and MCID-specific disease conditions, including OA.

Research usability. More research is required to determine the responsiveness to change, MDCs, and MCIDs in people with lower extremity OA and other populations. Administrative and respondent burden does not limit research use.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Dreinhofer K, Stucki G, Ewert T, Huber E, Ebenbichler G, Gutenbrunner C, et al. ICF Core Sets for osteoarthritis. J Rehabil Med 2004;44 Suppl:75–80.
2. Hayes KW, Johnson ME. Measures of adult general performance tests: the Berg Balance Scale, Dynamic Gait Index (DGI), Gait Velocity, Physical Performance Test (PPT), Timed Chair Stand Test, Timed Up and Go, and Tinetti Performance-Oriented Mobility Assessment (POMA). Arthritis Rheum 2003;49 Suppl:S28–42.
3. World Health Organization. International Classification of Functioning, Disability and Health. Geneva, Switzerland: WHO; 2001.
4. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis: consensus development at OMERACT III. J Rheumatol 1997;24:799–802.
5. Terwee CB, Mokkink LB, Steultjens MP, Dekker J. Performance-based methods for measuring the physical function of patients with osteoarthritis of the hip or knee: a systematic review of measurement properties. Rheumatology (Oxford) 2006;45:890–902.
6. Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. Scand J Med Sci Sports 2001;11:280–6.
7. Marks R. Reliability and validity of self-paced walking time measures for knee osteoarthritis. Arthritis Care Res 1994;7:50–3.
8. Gill S, McBurney H. Reliability of performance-based measures in people awaiting joint replacement surgery of the hip or knee. Physiother Res Int 2008;13:141–52.

9. Fransen M, Crosbie J, Edmonds J. Reliability of gait measurements in people with osteoarthritis of the knee. Phys Ther 1997;77:944–53.
10. Schilke JM, Johnson GO, Housh TJ, O'Dell JR. Effects of muscle-strength training on the functional status of patients with osteoarthritis of the knee joint. Nurs Res 1996;45:68–72.
11. Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskelet Disord 2005;6:3.
12. Wright AA, Cook CE, Baxter GD, Garcia J, Abbott JH. Relationship between the Western Ontario and McMaster Universities Osteoarthritis Index Physical Function Subscale and physical performance measures in patients with hip osteoarthritis. Arch Phys Med Rehabil 2010;91:1558–64.
13. Wright AA, Cook CE, Baxter GD, Dockerty JD, Abbott JH. A comparison of 3 methodological approaches to defining major clinically important improvement of 4 performance measures in patients with hip osteoarthritis. J Orthop Sports Phys Ther 2011;41:319–27.
14. Cibulka MT, White DM, Woehrle J, Harris-Hayes M, Enseki K, Fagerson TL, et al. Hip pain and mobility deficits. Hip osteoarthritis: clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopaedic section of the American Physical Therapy Association. J Orthop Sports Phys Ther 2009;39:A1–25.
15. Galea MP, Levinger P, Lythgo N, Cimoli C, Weller R, Tully E, et al. A targeted home- and center-based exercise program for people after total hip replacement: a randomized clinical trial. Arch Phys Med Rehabil 2008;89:1442–7.
16. Silva LE, Valim V, Pessanha AP, Oliveira LM, Myamoto S, Jones A, et al. Hydrotherapy versus conventional land-based exercise for the management of patients with osteoarthritis of the knee: a randomized clinical trial. Phys Ther 2008;88:12–21.
17. Fisher NM, Gresham GE, Abrams M, Hicks J, Horrigan D, Pendergast DR. Quantitative effects of physical therapy on muscular and functional performance in subjects with osteoarthritis of the knees. Arch Phys Med Rehabil 1993;74:840–7.
18. Altman RD, Moskowitz R. Intraarticular sodium hyaluronate (Hyalgan) in the treatment of patients with osteoarthritis of the knee: a randomized clinical trial. J Rheumatol 1998;25:2203–12.
19. Altman RD, Rosen JE, Bloch DA, Hatoum HT, Korner P. A double-blind, randomized, saline-controlled study of the efficacy and safety of EUFLEXXA for treatment of painful osteoarthritis of the knee, with an open-label safety extension (the FLEXX trial). Semin Arthritis Rheum 2009;39:1–9.
20. Kauppila AM, Kyllonen E, Mikkonen P, Ohtonen P, Laine V, Siira P, et al. Disability in end-stage knee osteoarthritis. Disabil Rehabil 2009; 31:370–80.
21. Pua YH, Clark RA, Bryant AL. Physical function in hip osteoarthritis: relationship to isometric knee extensor steadiness. Arch Phys Med Rehabil 2010;91:1110–6.
22. Thomas SG, Pagura SM, Kennedy D. Physical activity and its relationship to physical performance in patients with end stage knee osteoarthritis. J Orthop Sports Phys Ther 2003;33:745–54.
23. Cecchi F, Molino-Lova R, Di Iorio A, Conti AA, Mannoni A, Lauretani F, et al. Measures of physical performance capture the excess disability associated with hip pain or knee pain in older persons. J Gerontol A Biol Sci Med Sci 2009;64:1316–24.
24. McCarthy CJ, Oldham JA. The reliability, validity and responsiveness of an aggregated locomotor function (ALF) score in patients with osteoarthritis of the knee. Rheumatology (Oxford) 2004;43:514–7.
25. Rejeski WJ, Ettinger WH Jr, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995;3:157–67.
26. Shields RK, Enloe LJ, Evans RE, Smith KB, Steckel SD. Reliability, validity, and responsiveness of functional tests in patients with total joint replacement. Phys Ther 1995;75:169–79.
27. Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 2006;86:1489–96.
28. Steultjens MP, Roorda LD, Dekker J, Bijlsma JW. Responsiveness of observational and self-report methods for assessing disability in mobility in patients with osteoarthritis. Arthritis Rheum 2001;45:56–61.
29. Juhakoski R, Tenhonen S, Anttonen T, Kauppinen T, Arokoski JP. Factors affecting self-reported pain and physical function in patients with hip osteoarthritis. Arch Phys Med Rehabil 2008;89:1066–73.
30. Bohannon RW. Comfortable and maximum walking speed of adults aged 20–79 years: reference values and determinants. Age Ageing 1997;26:15–9.
31. Cesari M, Kritchevsky SB, Penninx BW, Nicklas BJ, Simonsick EM, Newman AB, et al. Prognostic value of usual gait speed in well-functioning older people: results from the Health, Aging and Body Composition Study. J Am Geriatr Soc 2005;53:1675–80.
32. Verghese J, Holtzer R, Lipton RB, Wang C. quantitative gait markers and incident fall risk in older adults. J Gerontol A Biol Sci Med Sci 2009;64A:896–901.
33. Stratford PW, Kennedy D, Pagura SM, Gollish JD. The relationship between self-report and performance-related measures: questioning the content validity of timed tests. Arthritis Rheum 2003;49:535–40.
34. Escalante A, Lichtenstein MJ, Hazuda HP. Walking velocity in aged persons: its association with lower extremity joint range of motion. Arthritis Rheum 2001;45:287–94.
35. Guralnik JM, Ferrucci L, Simonsick EM, Salive ME, Wallace RB. Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability. N Engl J Med 1995;332:556–61.
36. Grace EM, Gerecz EM, Kassam YB, Buchanan HM, Buchanan WW, Tugwell PS. 50-foot walking time: a critical assessment of an outcome measure in clinical therapeutic trials of antirheumatic drugs. Br J Rheumatol 1988;27:372–4.
37. Ward MM. Clinical measures in rheumatoid arthritis: which are most useful in assessing patients? J Rheumatol 1994;21:17–27.
38. Perera S, Mody SH, Woodman RC, Studenski SA. Meaningful change and responsiveness in common physical performance measures in older adults. J Am Geriatr Soc 2006;54:743–9.
39. Ostchega Y, Harris TB, Hirsch R, Parsons VL, Kington R. The prevalence of functional limitations and disability in older persons in the US: data from the National Health and Nutrition Examination Survey III. J Am Geriatr Soc 2000;48:1132–5.
40. Steffen TM, Hacker TA, Mollinger L. Age- and gender-related test performance in community-dwelling elderly people: Six-Minute Walk Test, Berg Balance Scale, Timed Up & Go Test, and gait speeds. Phys Ther 2002;82:128–37.
41. Cress ME, Schechtman KB, Mulrow CD, Fiatarone MA, Gerety MB, Buchner DM. Relationship between physical performance and self-perceived physical function. J Am Geriatr Soc 1995;43:93–101.
42. Kennedy D, Stratford PW, Pagura SM, Walsh M, Woodhouse LJ. Comparison of gender and group differences in self-report and physical performance measures in total hip and knee arthroplasty candidates. J Arthroplasty 2002;17:70–7.
43. Maly MR, Costigan PA, Olney SJ. Determinants of self-report outcome measures in people with knee osteoarthritis. Arch Phys Med Rehabil 2006;87:96–104.
44. Walsh M. Functional outcome measures: individuals one year post total knee arthroplasty versus healthy controls. Toronto (Ontario, Canada): University of Toronto; 1995.
45. Walsh M, Woodhouse LJ, Thomas SG, Finch E. Physical impairments and functional limitations: a comparison of individuals 1 year after total knee arthroplasty with control subjects. Phys Ther 1998;78: 248–58.
46. Finch E, Walsh M, Thomas SG, Woodhouse LJ. Functional ability perceived by individuals following total knee arthroplasty compared to age-matched individuals without knee disability. J Orthop Sports Phys Ther 1998;27:255–63.
47. Parent E, Moffet H. Comparative responsiveness of locomotor tests and questionnaires used to follow early recovery after total knee arthroplasty. Arch Phys Med Rehabil 2002;83:70–80.
48. Kreibich DN, Vaz M, Bourne RB, Rorabeck CH, Kim P, Hardie R, et al. What is the best way of assessing outcome after total knee replacement? Clin Orthop Relat Res 1996;331:221–5.
49. Madsen OR, Brot C. Assessment of extensor and flexor strength in the individual gonarthrotic patient: interpretation of performance changes. Clin Rheumatol 1996;15:154–60.
50. Logerstedt DS, Snyder-Mackler L, Ritter RC, Axe MJ. Knee pain and mobility impairments: meniscal and articular cartilage lesions. J Orthop Sports Phys Ther 2010;40:A1–35.
51. Farquhar S, Snyder-Mackler L. The Chitranjan Ranawat Award: the nonoperated knee predicts function 3 years after unilateral total knee arthroplasty. Clin Orthop Relat Res 2010;468:37–44.
52. Floren M, Reichel H, Davis J, Laskin RS. The mini-incision mid-vastus approach for total knee arthroplasty. Oper Orthop Traumatol 2008;20: 534–43.
53. Zeni JA Jr, Snyder-Mackler L. Clinical outcomes after simultaneous bilateral total knee arthroplasty: comparison to unilateral total knee arthroplasty and healthy controls. J Arthroplasty 2010;25:541–6.
54. Zeni JA Jr, Snyder-Mackler L. Early postoperative measures predict 1- and 2-year outcomes after unilateral total knee arthroplasty: importance of contralateral limb strength. Phys Ther 2010;90:43–54.
55. Ettinger WH Jr, Burns R, Messier SP, Applegate W, Rejeski WJ, Morgan T, et al. A randomized trial comparing aerobic exercise and resistance exercise with a health education program in older adults with knee osteoarthritis: the Fitness Arthritis and Seniors Trial (FAST). JAMA 1997;277:25–31.
56. Fransen M, Nairn L, Winstanley J, Lam P, Edmonds J. Physical activity for osteoarthritis management: a randomized controlled clinical trial evaluating hydrotherapy or Tai Chi classes. Arthritis Rheum 2007;57: 407–14.
57. Bennell KL, Hunt MA, Wrigley TV, Hunter DJ, McManus FJ, Hodges

PW, et al. Hip strengthening reduces symptoms but not knee load in people with medial knee osteoarthritis and varus malalignment: a randomised controlled trial. Osteoarthritis Cartilage 2010;18:621–8.

58. McKnight PE, Kasle S, Going S, Villanueva I, Cornett M, Farr J, et al. A comparison of strength training, self-management, and the combination for early osteoarthritis of the knee. Arthritis Care Res (Hoboken) 2010;62:45–53.

59. Talbot LA, Gaines JM, Huynh TN, Metter EJ. A home-based pedometer-driven walking program to increase physical activity in older adults with osteoarthritis of the knee: a preliminary study. J Am Geriatr Soc 2003;51:387–92.

60. Clarke AK. A double-blind comparison of naproxen against indometacin in osteoarthrosis. Arzneimittelforschung 1975;25:302–4.

61. Singh JA, O'Byrne M, Harmsen S, Lewallen D. Predictors of moderate-severe functional limitation after primary total knee arthroplasty (TKA): 4701 TKAs at 2-years and 2935 TKAs at 5-years. Osteoarthritis Cartilage 2010;18:515–21.

62. Zeni JA Jr, Axe MJ, Snyder-Mackler L. Clinical predictors of elective total joint replacement in persons with end-stage knee osteoarthritis. BMC Musculoskelet Disord 2010;11:86.

63. Zeni JA Jr, Snyder-Mackler L. Preoperative predictors of persistent impairments during stair ascent and descent after total knee arthroplasty. J Bone Joint Surg Am 2010;92:1130–6.

64. Sevick MA, Bradham DD, Muender M, Chen GJ, Enarson C, Dailey M, et al. Cost-effectiveness of aerobic and resistance exercise in seniors with knee osteoarthritis. Med Sci Sports Exerc 2000;32:1534–40.

65. Almeida GJ, Schroeder CA, Gil AB, Fitzgerald GK, Piva SR. Interrater reliability and validity of the stair ascend/descend test in subjects with total knee arthroplasty. Arch Phys Med Rehabil 2010;91:932–8.

66. Collins E, O'Connell S, Jelinek C, Miskevics S, Budiman-Mak E. Evaluation of psychometric properties of Walking Impairment Questionnaire in overweight patients with osteoarthritis of knee. J Rehabil Res Dev 2008;45:559–66.

67. Mizner RL, Petterson SC, Clements KE, Zeni JA Jr, Irrgang JJ, Snyder-Mackler L. Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments: a longitudinal analysis of outcomes. J Arthroplasty 2011;26:728–37.

68. Davis MA, Ettinger WH, Neuhaus JM, Mallon KP. Knee osteoarthritis and physical functioning: evidence from the NHANES-I Epidemiologic Followup Study. J Rheumatol 1991;18:591–8.

69. Stratford PW, Kennedy DM, Riddle DL. New study design evaluated the validity of measures to assess change after hip or knee arthroplasty. J Clin Epidemiol 2009;62:347–52.

70. Beaton D, Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in patients with advanced osteoarthritis of the hip or knee: invited commentary. Phys Ther 2006;86:1496–500.

71. Balke B. A simple field test for the assessment of physical fitness: rep 63-6. Rep Civ Aeromed Res Inst US 1963:1–8.

72. McGavin CR, Gupta SP, McHardy GJ. Twelve-minute walking test for assessing disability in chronic bronchitis. Br Med J 1976;1:822–3.

73. Butland RJ, Pang J, Gross ER, Woodcock AA, Geddes DM. Two-, six-, and 12-minute walking tests in respiratory disease. Br Med J (Clin Res Ed) 1982;284:1607–8.

74. ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories. ATS statement: guidelines for the six-minute walk test. Am J Respir Crit Care Med 2002;166:111–7.

75. Stratford PW, Kennedy DM, Maly MR, Macintyre NJ. Quantifying self-report measures' overestimation of mobility scores postarthroplasty. Phys Ther 2010;90:1288–96.

76. Foley A, Halbert J, Hewitt T, Crotty M. Does hydrotherapy improve strength and physical function in patients with osteoarthritis: a randomised controlled trial comparing a gym based and a hydrotherapy based strengthening programme. Ann Rheum Dis 2003;62:1162–7.

77. Deyle GD, Henderson NE, Matekel RL, Ryder MG, Garber MB, Allison SC. Effectiveness of manual physical therapy and exercise in osteoarthritis of the knee: a randomized, controlled trial. Ann Intern Med 2000;132:173–81.

78. Moffet H, Collet JP, Shapiro SH, Paradis G, Marquis F, Roy L. Effectiveness of intensive rehabilitation on functional ability and quality of life after first total knee arthroplasty: a single-blind randomized controlled trial. Arch Phys Med Rehabil 2004;85:546–56.

79. Frestedt JL, Kuskowski MA, Zenk JL. A natural seaweed derived mineral supplement (Aquamin F) for knee osteoarthritis: a randomised, placebo controlled pilot study. Nutr J 2009;8:7.

80. Maly MR, Costigan PA, Olney SJ. Role of knee kinematics and kinetics on performance and disability in people with medial compartment knee osteoarthritis. Clin Biomech (Bristol, Avon) 2006;21:1051–9.

81. Crosbie J, Naylor J, Harmer A, Russell T. Predictors of functional ambulation and patient perception following total knee replacement and short-term rehabilitation. Disabil Rehabil 2010;32:1088–98.

82. Parent E, Moffet H. Preoperative predictors of locomotor ability two months after total knee arthroplasty for severe osteoarthritis. Arthritis Rheum 2003;49:36–50.

83. Lord SR, Menz HB. Physiologic, psychologic, and health predictors of 6-minute walk performance in older people. Arch Phys Med Rehabil 2002;83:907–11.

84. Stratford PW, Kennedy DM. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. J Clin Epidemiol 2006;59:160–7.

85. French HP, Fitzpatrick M, FitzGerald O. Responsiveness of physical function outcomes following physiotherapy intervention for osteoarthritis of the knee: an outcome comparison study. Physiotherapy. In press.

86. Rikli RE, Jones CJ. The development and validation of a functional fitness test for community-residing older adults. J Aging Phys Act 1999;7:129–61.

87. Rikli RE, Jones CJ. Functional fitness normative scores for community-residing older adults, ages 60–94. J Aging Phys Act 1999;7:162–81.

88. Salzman SH. The 6-min walk test: clinical and research role, technique, coding, and reimbursement. Chest 2009;135:1345–52.

89. Jenkins S, Cecins N, Camarri B, Williams C, Thompson P, Eastwood P. Regression equations to predict 6-minute walk distance in middle-aged and elderly adults. Physiother Theory Pract 2009;25:516–22.

90. Price AJ, Webb J, Topf H, Dodd CA, Goodfellow JW, Murray DW, et al. Rapid recovery after oxford unicompartmental arthroplasty through a short incision. J Arthroplasty 2001;16:970–6.

91. Hamilton DM, Haennel RG. Validity and reliability of the 6-minute walk test in a cardiac rehabilitation population. J Cardiopulm Rehabil 2000;20:156–64.

92. Redelmeier DA, Bayoumi AM, Goldstein RS, Guyatt GH. Interpreting small differences in functional status: the Six Minute Walk test in chronic lung disease patients. Am J Respir Crit Care Med 1997;155:1278–82.

93. Schoindre Y, Meune C, Dinh-Xuan AT, Avouac J, Kahan A, Allanore Y. Lack of specificity of the 6-minute walk test as an outcome measure for patients with systemic sclerosis. J Rheumatol 2009;36:1481–5.

94. Harada ND, Chiu V, Stewart AL. Mobility-related function in older adults: assessment with a 6-minute walk test. Arch Phys Med Rehabil 1999;80:837–41.

95. Gibbons WJ, Fruchter N, Sloan S, Levy RD. Reference values for a multiple repetition 6-minute walk test in healthy adults older than 20 years. J Cardiopulm Rehabil 2001;21:87–93.

96. Csuka M, McCarty DJ. Simple method for measurement of lower extremity muscle strength. Am J Med 1985;78:77–81.

97. Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine (Phila Pa 1976) 1998;23:2412–21.

98. Jones CJ, Rikli RE, Beam WC. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. Res Q Exerc Sport 1999;70:113–9.

99. Boonstra MC, De Waal Malefijt MC, Verdonschot N. How to quantify knee function after total knee arthroplasty? Knee 2008;15:390–5.

100. Catani F, Innocenti B, Belvedere C, Labey L, Ensini A, Leardini A. The Mark Coventry Award: articular contact estimation in TKA using in vivo kinematics and finite element analysis. Clin Orthop Relat Res 2010;468:19–28.

101. Arnold CM, Faulkner RA. The effect of aquatic exercise and education on lowering fall risk in older adults with hip osteoarthritis. J Aging Phys Act 2010;18:245–60.

102. Piva SR, Gil AB, Almeida GJ, DiGioia AM 3rd, Levison TJ, Fitzgerald GK. A balance exercise program appears to improve function for patients with total knee arthroplasty: a randomized clinical trial. Phys Ther 2010;90:880–94.

103. Wang C, Schmid CH, Hibberd PL, Kalish R, Roubenoff R, Rones R, et al. Tai Chi is effective in treating knee osteoarthritis: a randomized controlled trial. Arthritis Rheum 2009;61:1545–53.

104. Fujita T, Fujii Y, Munezane H, Ohue M, Takagi Y. Analgesic effect of raloxifene on back and knee pain in postmenopausal women with osteoporosis and/or osteoarthritis. J Bone Miner Metab 2010;28:477–84.

105. Wright AA, Hegedus EJ, David Baxter G, Abbott JH. Measurement of function in hip osteoarthritis: developing a standardized approach for physical performance measures. Physiother Theory Pract 2011;27:253–62.

106. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994;49:M85–94.

107. Jette AM, Jette DU, Ng J, Plotkin DJ, Bach MA. Are performance-based measures sufficiently reliable for use in multicenter trials? J Gerontol A Biol Sci Med Sci 1999;54:M3–6.

108. Lipsitz LA, Jonsson PV, Kelley MM, Koestner JS. Causes and corre-

lates of recurrent falls in ambulatory frail elderly. J Gerontol 1991;46: M114–22.

109. Mathias S, Nayak US, Isaacs B. Balance in elderly patients: the "get-up and go" test. Arch Phys Med Rehabil 1986;67:387–9.

110. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991; 39:142–8.

111. Hinman RS, Bennell KL, Crossley KM, McConnell J. Immediate effects of adhesive tape on pain and disability in individuals with knee osteoarthritis. Rheumatology (Oxford) 2003;42:865–9.

112. Kennedy DM, Hanna SE, Stratford PW, Wessel J, Gollish JD. Preoperative function and gender predict pattern of functional recovery after hip and knee arthroplasty. J Arthroplasty 2006;21:559–66.

113. Kraemer WJ, Ratamess NA, Anderson JM, Maresh CM, Tiberio DP, Joyce ME, et al. Effect of a cetylated fatty acid topical cream on functional mobility and quality of life of patients with osteoarthritis. J Rheumatol 2004;31:767–74.

114. Arnold CM, Faulkner RA. The history of falls and the association of the timed up and go test to falls and near-falls in older adults with hip osteoarthritis. BMC Geriatr 2007;7:17.

115. Halket A, Stratford PW, Kennedy DM, Woodhouse LJ. Using hierarchical linear modeling to explore predictors of pain after total hip and knee arthroplasty as a consequence of osteoarthritis. J Arthroplasty 2010;25:254–62.

116. Bohannon RW. Reference values for the timed up and go test: a descriptive meta-analysis. J Geriatr Phys Ther 2006;29:64–8.

117. Shumway-Cook A, Brauer S, Woollacott M. Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go Test. Phys Ther 2000;80:896–903.

118. Salarian A, Horak FB, Zampieri C, Carlson-Kuhta P, Nutt JG, Aminian K. iTUG, a sensitive and reliable measure of mobility. IEEE Trans Neural Syst Rehabil Eng 2010;18:303–10.

119. Lin MR, Hwang HF, Hu MH, Wu HD, Wang YW, Huang FC. Psychometric comparisons of the timed up and go, one-leg stand, functional reach, and Tinetti balance measures in community-dwelling older people. J Am Geriatr Soc 2004;52:1343–8.

120. Ferreira PH, Ferreira ML, Maher CG, Herbert RD, Refshauge K. Specific stabilisation exercise for spinal and pelvic pain: a systematic review. Aust J Physiother 2006;52:79–88.

121. De Morton NA, Keating JL, Jeffs K. Exercise for acutely hospitalised older medical patients. Cochrane Database Syst Rev 2007;1: CD005955.

122. Noren AM, Bogren U, Bolin J, Stenstrom C. Balance assessment in patients with peripheral arthritis: applicability and reliability of some clinical assessments. Physiother Res Int 2001;6:193–204.

123. Maly MR, Costigan PA, Olney SJ. Contribution of psychosocial and mechanical variables to physical performance measures in knee osteoarthritis. Phys Ther 2005;85:1318–28.

124. Sasaki K, Senda M, Nishida K, Ota H. Preoperative time required for the timed "up and go" test in women with hip osteoarthritis could predict a deep venous thrombosis complication after total hip arthroplasty. Acta Med Okayama 2010;64:197–201.

125. McMeeken J, Stillman B, Story I, Kent P, Smith J. The effects of knee extensor and flexor muscle training on the timed-up-and-go test in individuals with rheumatoid arthritis. Physiother Res Int 1999;4: 55–67.

126. Hinman RS, Heywood SE, Day AR. Aquatic physical therapy for hip and knee osteoarthritis: results of a single-blind randomized controlled trial. Phys Ther 2007;87:32–43.

127. Rockwood K, Awalt E, Carver D, MacKnight C. Feasibility and measurement properties of the functional reach and the timed up and go tests in the Canadian study of health and aging. J Gerontol A Biol Sci Med Sci 2000;55:M70–3.

128. Strand LI, Moe-Nilssen R, Ljunggren AE. Back Performance Scale for the assessment of mobility-related activities in people with back pain. Phys Ther 2002;82:1213–23.

129. Strand LI, Wie SL. The Sock Test for evaluating activity limitation in patients with musculoskeletal pain. Phys Ther 1999;79:136–45.

130. Bieleman HJ, Reneman MF, van Ittersum MW, van der Schans CP, Groothoff JW, Oosterveld FG. Self-reported functional status as predictor of observed functional capacity in subjects with early osteoarthritis of the hip and knee: a diagnostic study in the CHECK cohort. J Occup Rehabil 2009;19:345–53.

131. Bohannon RW. Reference values for the five-repetition sit-to-stand test: a descriptive meta-analysis of data from elders. Percept Mot Skills 2006;103:215–22.

Summary Table for Clinical Physical Performance Measures of Activity*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
SPWT	Assesses lower extremity function through timed walking over short distances (usually <50 meters)	Performance based	Minimal; <5 min	Minimal; <5 min Easy to set up with minimal equipment Hand score at completion of test Usually converted to speed (e.g., meters/min, meters/sec)	Gait speeds <1 meter/sec associated with high risk of dysfunction in older people (31) Normative values available (30)	Test–retest $ICC_{2,1}$ 0.91 (95% CI 0.81, 0.97) (11) Intertester $ICC_{1,1}$ 0.96 (95% CI 0.93, 0.98) (8) 40-meter SPWT: SEM 1.32 sec (11) 40-meter SPWT: MCD_{90} 4.04 sec (11)	Construct: r = 0.66 with Index of Severity for Knee (7), r = 0.59 with Lower Extremity Functional Scale (33)	SRM 0.79 (95% CI 0.66, 1.45; n = 89) after THR or TKR (11) 40-meter SPWT: MCID 0.2–0.3 meter/sec (13)	Measures a core activity limitation in lower extremity OA Stable over short periods Responsive following rehabilitation and surgery Easy to conduct Acceptable	Can be affected by practice effects, depression, and cognitive status
SCT	Assesses lower extremity strength, power, and times (sec) Ability to negotiate a defined number of stairs	Performance based	Minimal; <5 min	Minimal; <5 min Easy to set up provided access to stairs Hand score on completion of test	No relevant information found	Test–retest $ICC_{2,1}$ 0.90 (95% CI 0.79, 0.96) (11) Intertester $ICC_{2,1}$ 0.94 (95% CI 0.55, 0.98) post-TKR (65) 9-step SCT: MCD_{90} 5.5 sec (11)	Construct function: r = −0.53 with WOMAC-PF (6) Construct strength: r = −0.50 to 0.52 with quadriceps and hamstrings (43)	SRM 1.98 (95% CI 1.68, 2.42; n = 89) after THR or TKR (11) ES 0.84 12 months after TKR (67)	Measures a core activity limitation in lower extremity OA Stable over short periods Responsive following rehabilitation and surgery Easy to conduct Acceptable	Step number, step height, and use of hand rail can vary performance and need to be kept consistent on multiple assessments
6MWT	Assesses endurance and the ability to walk over longer distances Measures distance covered in 6 min	Performance based	<10 min	<10 min Easy to set up, provided sufficient space, with minimal equipment Hand score on completion of test	Distances associated with age, sex, height, and in women, BMI Predictive regression equations available for older adults (89)	Test–retest $ICC_{2,1}$ 0.94 (95% CI 0.88, 0.98) (11) SEM 26.9 meters (95% CI 21.1, 34.8) (11) MCD_{90} 61.3 meters (11)	Construct endurance: r = 0.71 with Vo_{2max} (90) Construct function: R = 0.62 with SF-36 function (91) Predictive: R^2 = 0.66 for walking ability after TKR (47)	SRM 1.90 (95% CI 1.46, 2.39; n = 61) after THR or TKR (11) ES 0.43, SRM 0.54 following physical therapy (85) MCID 50–54 meters (38,92)	Measures a common limitation in lower extremity OA Stable over short periods Responsive following rehabilitation and surgery Appropriate for early and end-stage OA	Use of encouragement and the number of trials performed can vary performance Contraindicated in unstable coronary disease Precautions available
CST	Assesses lower extremity strength and power and the ability (number) to rise and sit in a chair over 30 sec or the time taken to do a set number of repetitions	Performance based	Minimal; <3 min	<3 min Easy to set up with minimal equipment and score on completion of test	Inability to rise without using arms is a predictor of falls in elderly (OR 3.4; 95% CI 1.2, 9.4) (108) Age and sex normative values available (86,131)	Test–retest intrasession $ICC_{1,1}$ 0.95 (95% CI 0.93, 0.97) (8) 2–5-day ICC 0.95 (95% CI 0.93, 0.97) (98) Intertester $ICC_{1,1}$ 0.93 (95% CI 0.87, 0.96) (8) SEM 0.7 stands (8) MCD_{90} 1.6 stands (8)	Construct function: r = 0.66 with walking speed (107) Construct strength: r = 0.71–0.78 with the leg press test for elderly (98)	ES 0.36, SRM 0.39 following physical therapy (85) 30-sec CST: MCID 2–2.6 reps (13)	Measures a common activity limitation for lower extremity OA Easy to conduct	A practice trial is necessary to minimize baseline practice effects Pain may limit acceptability and result in potential floor effects
TUG	Assesses lower extremity strength and power and basic mobility skills Measures time (sec) to rise from a chair, walk 3 meters, turn, and walk back to sit in chair	Performance based	Minimal; <3 min	<3 min Easy to set up with minimal equipment Hand score on completion of test	TUG >10 sec is a predictor of near falls in hip OA (OR 3.1) (114) Older adults with TUG >14 sec have higher risk for falls (117) Normative values available (116)	Test–retest intrasession $ICC_{2,1}$ 0.95 (95% CI 0.95, 0.97) in elderly (40) 3–6 month $ICC_{2,1}$ 0.75 (95% CI 0.51, 0.98) (11) SEM 1.07 sec (95% CI 0.86, 1.41) (8) MCD_{90} 2.49 sec (8)	Construct function: r = −0.51 with Bartel Index (110) Construct strength: r = 0.49–0.51 with quadriceps and hamstrings (123) Predictive: TUG ≥5.3 83% sensitivity, 61% specificity to predict DVT post-THR (124)	SRM 1.04 (95% CI 0.89, 1.61) after THR or TKR (11) ES 0.33, SRM 0.35 after physical therapy (85) MCID reduction of 0.8–1.4 sec (13)	Measures a core activity for lower extremity OA Responsive following rehabilitation and surgery Easy to conduct Acceptable	Lack of stability when reassessed over longer (>6 mos) intervals Can be affected by ceiling and floor effects Multiple subcomponents may limit interpretation
Sock Test	Assesses lower extremity flexibility and ability to put on footwear Measured on 0–3 ordinal scale of ability and ease	Performance based	Minimal; <5 min	Minimal; <5 min Easy to set up with minimal equipment Hand score (grade) on completion of test	No relevant information found	Intertester weighted κ = 0.79 (95% CI 0.5, 1.0) (129)	Construct function: ρ = 0.45 with DRI (130) Predictive: 6–12 times increased likelihood of dysfunction with scores of 2 and 3 (129)	No relevant evidence found	Direct measure of a common activity limitation in lower extremity OA Easy to conduct Acceptable	Scores can be influenced by age and BMI and need to be considered when interpreting results
LCT	Assesses the ability to lift up, carry, and replace a weight over a short distance Measures the time (sec) taken and ease	Performance based	Minimal; <5 min	Minimal; <5 min Hand score on completion of test	No relevant information found	14-day test–retest: r = 0.92 (25) 3-month test–retest: r = 0.77 (25) SEM 0.49–0.50 sec in knee OA (25)	Construct function: r = −0.40 with treadmill time (25), r = −0.38 with Vo_{2peak} (25), r = 0.34 with FAST Functional Activities Inventory (25)	$P < 0.001$ following aerobic exercise in knee OA (55) $P = 0.003$ following resistive exercise program in knee OA (55)	Direct measure of a common activity limitation in lower extremity OA Appears stable over short periods	Small systematic improvements over time may reflect motivational or learning effects Stability may be influenced by external factors over longer periods Requires a contextual setup
Car Task	Assesses the ability to get in and out of a car Measures the time (sec) taken and ease	Performance based	<10 min	<10 min Hand score on completion of test	No relevant information found	14-day test–retest: r = 0.88 (25) 3-month test–retest: r = 0.86 (25) SEM 0.88–0.97 sec in knee OA (25)	Construct function: r = −0.45 with treadmill time (25), r = 0.35–0.38 with FAST Functional Activities Inventory (25) Construct strength: r = −0.46 with knee strength (25)	$P < 0.001$ following aerobic exercise in knee OA (55) $P = 0.003$ following resistive exercise program in knee OA (55)	Direct measure of a common activity limitation in lower extremity OA Appears stable over short periods	May be influenced by learning effects on repeat testing occasions Requires access to a car for testing

* SPWT = Self-Paced Walk Test; ICC = intraclass correlation coefficient; 95% CI = 95% confidence interval; SEM = standard error of measurement; MCD_{90} = minimal detectable change at 90% confidence; r = Pearson's correlation coefficient; SRM = standardized response mean; THR = total hip replacement; TKR = total knee replacement; MCID = minimum clinically important difference; OA = osteoarthritis; SCT = Stair Climb Test; WOMAC-PF = Western Ontario and McMaster Universities Osteoarthritis Index physical functioning subscale; ES = effect size; 6MWT = Six-Minute Walk Test; BMI = body mass index; Vo_{2max} = maximum oxygen consumption; SF-36 = Short Form 36; CST = Chair Stand Test; OR = odds ratio; TUG = Timed Up & Go; DVT = deep vein thrombosis; DRI = Disability Rating Index; LCT = Lift and Carry Test; FAST = Fitness and Activity Seniors Trial.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S473–S485
DOI 10.1002/acr.20567

Measures of Self-Efficacy

Arthritis Self-Efficacy Scale (ASES), Arthritis Self-Efficacy Scale-8 Item (ASES-8), Children's Arthritis Self-Efficacy Scale (CASE), Chronic Disease Self-Efficacy Scale (CDSES), Parent's Arthritis Self-Efficacy Scale (PASE), and Rheumatoid Arthritis Self-Efficacy Scale (RASE)

TERESA J. BRADY

INTRODUCTION

Enhancing self-efficacy has become an essential feature of many arthritis management interventions because of its robust relationships with health behaviors and health status. Empirical studies document that self-efficacy predicts health behaviors such as physical activity, eating behaviors, and pain coping strategies (1). In rheumatoid arthritis and osteoarthritis, self-efficacy has also been correlated with measures of health status such as daily pain and mood ratings (2), pain, stiffness, function, and physical and mental well-being (3); it has also been correlated with changes in pain, function, and depression (4). Adherence with medications and other health recommendations has also been associated with self-efficacy (5,6). In addition to evidence that self-efficacy is associated with health behaviors, current and future health status, and adherence to health recommendations, the fact that self-efficacy can change through efficacy-enhancing interventions makes it a rich target of arthritis interventions (1).

Self-efficacy, defined in Bandura's seminal 1977 article as "the conviction that one can successfully execute the behavior required to produce the outcomes" (7), was hypothesized to influence whether a behavior was initiated and sustained despite obstacles or adverse experiences, and to influence the level of effort invested in the behavior. Bandura's definition of self-efficacy evolved slightly over time; in his 1997 publication, Bandura defined self-efficacy as "belief in one's capability to organize and execute the courses of action required to produce given attainments" (8). Bandura has consistently described self-efficacy as domain specific and distinct from other constructs in social learning theory such as outcome expectations, defined as a person's estimate that a given behavior will lead to certain outcomes (7). Self-efficacy beliefs are also conceptualized as distinct from actual ability to perform a task (e.g., can I ride a bicycle), actual task performance (e.g., do I ride a bicycle), or intention to perform task (e.g., do I intend to ride a bicycle) (8,9). These different types of beliefs are clearly distinguished in Gecht et al's survey of exercise beliefs and habits among people with arthritis (10). In that survey, respondents were asked about their self-efficacy expectations regarding exercise ("If I want to exercise, I know I can do it"), and their outcome expectations regarding exercise ("regular exercise will probably make my arthritis worse in the future"); they were also asked to report their actual behavior (how often they did specific exercises in the past 2 weeks). Self-efficacy theory hypothesizes that both efficacy expectations and outcome expectations influence whether or not an individual will initiate and sustain a specific behavior (7). Gecht et al found that positive outcome expectations and self-efficacy for exercise were associated with participation in exercise (10). Conversely, self-efficacy theory predicts that if a patient believes that they can exercise (self-efficacy expectation) but also believes that exercise will be harmful for their arthritis (outcome expectation), the patient would be less likely to exercise than if they expected positive outcomes from exercise (7). Social learning theory suggests that it is important for clinicians and others hoping to help a person adopt health behaviors to understand both whether the person believes they can perform the behavior, and whether they believe that behavior will lead to positive outcomes.

Outcome expectations have received very little attention in arthritis-related research, but self-efficacy has been measured extensively (11). This review focuses on self-efficacy measures in the domain of arthritis management, and measures frequently used in arthritis management intervention studies (i.e., Arthritis Self-Efficacy Scales [12], Rheumatoid Arthritis Self-Efficacy Scale [13]). One nonarthritis specific measure, the Chronic Diseases Self-Efficacy Scale

The findings and conclusions herein are those of the author and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Teresa J. Brady, PhD: Centers for Disease Control and Prevention, Atlanta, Georgia.

Address correspondence to Teresa J. Brady, PhD, Centers for Disease Control and Prevention, 4770 Buford Highway NE, MS K-51, Atlanta, GA 30341. E-mail: Tob9@cdc.gov.

Submitted for publication March 29, 2011; accepted in revised form July 20, 2011.

(14), is included because it is frequently used in evaluation of self-management education programs. The review also includes a child-focused measure, the Children's Arthritis Self-Efficacy Scale (15), and a companion scale focused on parents' self-efficacy to manage arthritis-specific parenting tasks, the Parent's Arthritis Self-Efficacy Scale (16).

There are a number of related domain-specific measures, such as exercise self-efficacy scales (17,18), self-efficacy for managing anterior cruciate ligament injuries (19), and chronic pain self-efficacy scales (20–22), that are not reviewed here. However, all are included in a 2006 review by van Hartingsveld et al that covers a wide number of measures of patient expectations, including self-efficacy, across a wide range of musculoskeletal conditions (23).

Also not included in this review are scales which include the term self-efficacy in their titles, but do not measure domain- or behavior-specific efficacy beliefs such as the Generalized Self-Efficacy Scale (24,25) and the Self-Efficacy Scale (26). These general scales measure global beliefs in self-efficacy without specifying activities or conditions (e.g., "I can handle whatever comes my way," or "I can always manage to solve difficult problems") and are designed to assess a unidimensional global perception or static trait (25) rather than changeable domain- or behavior-specific beliefs. As such, these general scales appear more closely related to measures of perceived coping competence or mastery (27) rather than the domain-specific construct of self-efficacy as delineated by Bandura (7).

One area in the conceptualization and operationalization of self-efficacy that remains unclear in the literature is the delineation between efficacy expectations and outcome expectations, and this ambiguity is reflected in self-efficacy measures. In his more recent writings, Bandura has focused on self-efficacy as a person's belief in their "capability to produce given attainments" (8,18) rather than to execute the behavior required to produce outcomes (7); although in distinguishing self-efficacy from locus of control, Bandura described perceived self-efficacy as "beliefs about whether one can produce certain actions" while describing locus of control as "beliefs about whether actions affect outcomes" (8). Bandura cautioned against confusing performance, an accomplishment specified by descriptive markers (e.g., the academic grades of A, B, C, D, and F), with outcome, defined as something that flows from performance (specifically positive or negative physical, social, or self-evaluative effects) (8). This delineation between performance (attainment) and the physical, social, or self-evaluative effects (outcomes) that follow from that attainment is reasonably clear when considering academic grades (can I perform to the level of an A) but less clear when examining symptom relief, such as relief of pain (28). Some arthritis-related self-efficacy measures, such as the Arthritis Self-Efficacy Scales (12), consider symptom relief part of the efficacy belief (as a descriptive marker of the accomplishment), while others, such as the Rheumatoid Arthritis Self-Efficacy Scale (13), which focuses more on the execution of behaviors, consider symptom relief an outcome expectation (and not part of the efficacy belief). These distinctions are highlighted in the following review of measures.

ARTHRITIS SELF-EFFICACY SCALE (ASES)

Description

Purpose. The ASES was developed to measure patients' arthritis-specific self-efficacy, or patients' beliefs that they could perform specific tasks or behaviors to cope with the consequences of arthritis (12). The initial scale was published in 1989, and it was the first arthritis-specific measure of self-efficacy to appear in the literature. It remains the most widely used arthritis-specific measure. ASES was originally developed to help explain changes resulting from health education interventions in arthritis, and was developed using samples of people with arthritis attending community-based education programs, but has since been used in clinically based samples as well. The full 20-item scale has been translated into Swedish (29), Dutch (30), and Turkish (31). A shorter 8-item version of the ASES is available; see information of 8-item ASES reviewed elsewhere in this article.

Content. Items are designed to capture how certain the individual is that they can perform a specific activity or achieve a result. Items include specific behaviors (e.g., "Walk 100 feet on flat ground in 20 seconds" or "Scratch your upper back with both your right and left hands") and performance-attainment items (e.g., "Decrease your pain quite a bit," or "Control your fatigue").

Number of items. Original ASES has 20 items in 3 subscales: self-efficacy for managing pain (PSE), 5 items; self-efficacy for physical function (FSE), 9 items; and self-efficacy for controlling other symptoms (OSE), 6 items.

Response options/scale. Items are rated on a 10 (very uncertain) to 100 (very certain) rating scale, in 10-point increments. More recent versions of ASES have converted this to a 1–10 scale, as demonstrated on the web site where the scale is available (URL: http://patienteducation.stanford.edu/). The scale asks the respondent "how certain are you that you can" keep arthritis pain from interfering with your sleep (example from pain subscale), walk 10 steps downstairs in 10 seconds (example from function subscale), and control your fatigue, or do something to help yourself if you are feeling blue (examples from other symptoms subscale).

Recall period for items. Now or at the present time.

Endorsements. There are no noted endorsements.

Examples of use. ASES has been used in the evaluation of the Arthritis Self-Management Program (32,33) and in investigations of the association of self-efficacy to various health outcomes (34–36).

Practical Application

How to obtain. A copy of the ASES is included in the original publication (12). The full scale as well as the shortened scale are available from the web site of the Stanford Patient Education Research Center, URL: http://patienteducation.stanford.edu/.

Method of administration. Written, self-administered self-report questionnaire.

Scoring. Scoring instructions are provided on the web site (listed above), including instructions for handling missing values. Computer scoring is not required; scoring requires simple addition and division to calculate mean scores for each subscale.

Score interpretation. Score range is 10–100 or 1–10 for each subscale, depending on response options used. Higher scores indicate greater confidence or self-efficacy. No cut points or population-based norms are provided, although mean scores from the validation sample are provided.

Respondent burden. Time to complete is not reported, but estimated to be 5–10 minutes to complete all 20 items. Reading level appears appropriate, although some items may be complex (i.e., asking respondent to consider their certainty, about managing symptoms, to do desired activities).

Administrative burden. No training is required to administer ASES, scoring time requires calculation of 3 mean scores.

Translations/adaptations. The full 20-item ASES has been revalidated in Swedish (29,37), Dutch (30), and Turkish (31). The PSE and OSE subscales were evaluated for appropriateness for use with community-based samples in the UK, and determined to be appropriate, valid, and reliable with no modification necessary (38). The Swedish ASES was also adapted for chronic pain patients by replacing "arthritis" with "disease," and "arthritis pain" with "pain."

Psychometric Information

Method of development. A rheumatologist generated 23 original items; these items were refined, and an additional 20 items added following 3 focus groups of people with arthritis. Initial validation sample (n = 97) produced a 2-factor solution (other symptoms, function) using 25 items. The replication study (n = 144) produced a 3-factor solution (pain, other symptoms, function) using 20 items. Developers stated that the choice between the 2 and 3 factor solution was arbitrary, and they based the decision on the importance of pain and the performance of the other symptoms subscale in regard to depression. The Turkish translation of ASES resulted in a 4-factor solution, with function separated into upper and lower function. An item response theory analysis of the PSE and OSE subscales suggested the possibility that a single unidimensional factor underlies these 2 subscales, although 2 items needed to be removed to improve the fit of this 1-factor solution (38).

Acceptability. Original validation study does not address acceptability or missing items. The revalidation in the UK reported that no problems with comprehension, completion, or missing data were observed. It is not known if any ceiling or floor effects exist (28).

Reliability. Internal consistency reliability was estimated via Cronbach's alpha using data from 144 people who had registered for a community-based arthritis education program. Cronbach's alpha for PSE was 0.76, for FSE was 0.89, and for OSE was 0.87. Test–retest reliability (2–29 days between retesting) was calculated using respondents who had previously completed a community-based arthritis education program (n = 91). Test–retest reliability coefficients for PSE was 0.87, for FSE was 0.85, and for OSE was 0.90.

Validity. Initial validation was done using the same samples as were used for the development of the subscales, participants who had registered or completed a community-based arthritis education program. Construct validity was demonstrated by finding significant correlations among ASES subscales and measures of health status (pain, disability, and depression). Concurrent correlations between the pain subscale and health status measures were −0.29 for pain, −0.21 for disability, and −0.33 for depression. Concurrent correlations between the function subscale and health status measures were −0.29 for pain, −0.76 for disability, and −0.16 for depression. Concurrent correlations between other symptoms subscale and health status measures were −0.27 for pain, −0.25 for disability, and −0.44 for depression. Correlations of baseline self-efficacy with health status at 4 months for the pain subscale were −0.39 for pain, −0.21 for disability, and −0.45 for depression. Correlations of baseline self-efficacy with health status at 4 months for the function subscale were −0.30 for pain, −0.71 for disability, and −0.30 for depression. Correlations of baseline self-efficacy with health status at 4 months for the other symptoms subscale were −0.47 for pain, −0.21 for disability, and −0.59 for depression. In addition, participants who had participated in the arthritis education program showed greater change in self-efficacy scores than those that had not. Translated versions of the full ASES found similar theoretically relevant corrections with health status measures.

Ability to detect change. Sensitivity is unknown. Participants in the Arthritis Self-Management Program did demonstrate changes in ASES scores; although these changes were not statistically significant in the initial validation study, they were significant in subsequent evaluations of the Arthritis Self-Management Program.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Self-efficacy has become an important construct in understanding arthritis management and arthritis interventions, and the ASES (either the full 20-item scale or the 8-item scale) is the most common instrument used to measure self-efficacy for managing arthritis. The full ASES includes 3 subscales (pain, function, and other symptoms) assumed to be distinct in arthritis management, and this factor structure has been replicated in translated versions, although an item response theory analysis has raised questions about a single factor structure underlying PSE and OSE. The measure has been extensively used in evaluating education interventions and some clinical interventions as well. Items do cover a range of levels of task difficulty. All 3 subscales demonstrate good internal consistency and test–retest reliability, and are correlated with theoretically relevant health outcomes.

Caveats and cautions. In the initial validation article, the authors raise the question on whether they are capturing self-efficacy for behavior or outcome or some combination, but conclude that the distinction is not central for their purpose of identifying elements of health education programs that contribute to decreasing pain and increasing well-being and activity potential. Although the ASES does correlate with relevant health status measures, other aspects of self-efficacy theory, such as prediction of initiation or persistence of behavior, have not been examined. Some concerns have been raised that some items, particularly on the FSE, may tap actual performance rather than efficacy beliefs, and in the initial validation study, there was a 0.61 correlation between FSE and observed task performance (28). The majority of the validation studies have been conducted with community-based samples of people with arthritis; there has been no psychometric evaluation of comparability with a clinical population.

Clinical usability. Neither the administrative nor respondent burden should preclude its use, but the absence of any population-based norms or cut-off scores make it difficult to interpret an individual's score.

Research usability. The available psychometric data, including good reliability, validity, and demonstrated change with interventions, all suggest the ASES is appropriate for use in research. Neither the administrative nor respondent burden should preclude its use, although many investigators are using the 8-item version (reviewed elsewhere in this article).

ARTHRITIS SELF-EFFICACY SCALE-8 ITEM (ASES-8)

Description

Purpose. Because the label Arthritis Self-efficacy Scale is used in the literature to refer to both the full ASES (20 items, 3 subscales) and the shortened 8-item ASES, and because the reference for the validation of the original 20-item scale is used to support the 8-item version, investigators may be unaware of the extent of the psychometric support for the 8-item measure. For the purpose of this review this shorter measure has been christened the ASES-8. It is reviewed here as a separate measure because its psychometric support is significantly different than that of the original ASES. The ASES-8 was developed in the process of developing a set of Spanish-language health assessment instruments to be used in health promotion research. The original Spanish scale was published in 1995 (39). In 2003, a German language ASES-8 was validated with a small sample of patients with rheumatoid arthritis (n = 43) and 2 larger samples of people with fibromyalgia (40). While the ASES-8 is available in English, no psychometric studies of the English version have been published beyond mean and SD, and internal consistency reliability reported on the developer web site.

Content. The ASES-8 includes 2 items from the ASES pain subscale, 4 items from the ASES other symptoms subscale, and 2 new items related to preventing pain and fatigue from interfering with things you want to do.

Number of items. The total scale includes 8 items with no subscales.

Response options/scale. 1 (very uncertain) to 10 (very certain). Item stem for each question begins "How certain that you can. . . ."

Recall period for items. Now.

Endorsements. There are no noted endorsements.

Examples of use. ASES-8 has been used in evaluations of self-management education programs (41–43), physical activity interventions (44), and associations of self-efficacy with various health outcomes (45,46).

Practical Application

How to obtain. Spanish ASES-8 is available in original publication. Spanish and English versions are available from the web site of the Stanford Patient Education Research Center, URL: http://patienteducation.stanford.edu/.

Method of administration. The majority of the psychometric evaluation of the Spanish ASES-8 was done by interviewer administration; an undisclosed number were done by written self-report in the replication study.

Scoring. Scoring for the ASES-8 Spanish and English versions are available on the Stanford Patient Education Research Center web site, including instructions for handling missing items. No computer is necessary for scoring, which consists of calculating the mean of 8 item ratings.

Score interpretation. Scores range from 1–10. No cut points or population norms are available, although the mean scores for the Spanish validation sample, and an unpublished sample drawn from participants in the Arthritis Self-Management Program (n = 175), are provided on the Stanford Patient Education Research Center web site.

Respondent burden. Time to complete is not described, but assumed to be <5 minutes. Spanish items were pretested by interview and no difficulties noted; it is not clear if the written format is similarly problem free.

Administrative burden. No training is necessary for administration, and scoring is simple calculation of a mean score.

Translations/adaptations. ASES-8 was originally developed for Spanish-speaking respondents, although it is now being used with English respondents as well. The English version of the ASES-8 refers to both arthritis and fibromyalgia in each item. A German version was translated from the English version and tested in both rheumatoid arthritis and fibromyalgia samples; for the fibromyalgia sample, the term arthritis was replaced by fibromyalgia.

Psychometric Information

Method of development. The original item bank consisted of the 5-item pain self-efficacy and 6-item other symptom self-efficacy subscales of the original ASES, plus 2 other items "found to be useful in subsequent studies conducted by the investigators" (39). All items were translated into standard Spanish to avoid language variations found in various Spanish-speaking countries, using both back translation and translation by committee. Of the original 13 items, 5 were removed from the final scale because of low test–retest reliability, based on content review that suggested items were redundant or ambiguously worded, leaving 8 items for the short ASES in Spanish. Principal component factor analysis of the German version of the ASES-8 confirmed the single factor structure.

Acceptability. For the Spanish ASES-8, pretesting was done by administering the scale by interview and then questioning respondents about any difficulties they encountered; no difficulties were noted. It is not clear if missing data are common, or if there are ceiling or floor effects. The German version found no floor or ceiling effects.

Reliability. For the Spanish ASES-8, initial evaluation was by interview with respondents from 5 sites across the US and 1 site in Venezuela (n = 272); replication was done by interview and self-administered through the mail, with 151 subjects recruited from senior citizen centers and Hispanic service centers in the San Francisco Bay area. The Spanish ASES-8 had a Cronbach's alpha of 0.92, and item-to-scale correlations ranging from 0.65–0.83. Internal consistency reliability ranged from 0.88 for the Cuban-origin group to 0.93 for people from Mexican and Central American descent. Reliability was also high when looking exclusively at the self-administered subgroup in the replication study ($\alpha = 0.96$). Ten- to 14-day test–retest reliability was calculated using data from 25 participants in the replication study. Test–retest reliability was 0.69. The German translation of the English measure found reasonably similar reliability data ($\alpha = 0.90$), although the 8-week test–retest reliability was 0.51, using fibromyalgia and rheumatoid arthritis samples. The Stanford Patient Education Research Center web site, where the English ASES-8 is located, reports internal consistency reliability of 0.94 based on unpublished data from 175 participants in Stanford's Arthritis Self Management Program.

Validity. No validity data were presented for the Spanish ASES-8 or the English translation. The German translation of the English ASES-8 demonstrated theoretically relevant correlations between the ASES-8 and function (r = 0.20), depression (r = −0.53), and coping techniques (i.e., planning behavior; r = 0.35), and medium correlations with theoretically relevant constructs such as internal locus of control (r = 0.33), optimism (r = 0.39), and general self-efficacy (r = 0.40) among the 148 people with fibromyalgia in the initial evaluation.

Ability to detect change. Sensitivity of the Spanish or English ASES-8 is not reported, although it has been used in arthritis intervention studies and change has been reported. The German ASES documented medium size change (effect size 0.31) in a sample of 43 people with fibromyalgia in a clinical setting.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. A shorter version of the ASES has intuitive appeal for both research and clinical use. The German version, as used in people with fibromyalgia, has documented reliability, validity, and sensitivity to change. The Spanish ASES-8 has documented reliability, but validity is undocumented. The factor analysis of the German ASES-8 supports a single factor underlying responses. The English and Spanish versions have been used in intervention trials.

Caveats and cautions. A major weakness of the English version of the ASES-8 is the absence of any published psychometric data supporting its use. The articles documenting the development of the Spanish ASES-8 or the development of the full 20-item ASES are generally cited, yet neither of these articles describes any psychometric testing conducted on this 8-item measure. Further, 2 of the items are not part of the full ASES. Only the German version, tested primarily in people with fibromyalgia, has documented reliability, validity, sensitivity to change, and consistent underlying factor structure.

Clinical usability. Neither the administrative nor respondent burden should preclude its use, but psychometric information and population-based norms or cut-off scores will be necessary before it is useful in a clinical setting.

Research usability. The small administrative and respondent burden of the ASES-8 makes this an attractive option for research, but without psychometric evaluation of the English version of the ASES-8, it is difficult to determine its appropriateness for use in research with English-speaking subjects.

CHILDREN'S ARTHRITIS SELF-EFFICACY SCALE (CASE)

Description

Purpose. The CASE was designed to measure children's perceived ability to control or manage salient aspects of life with juvenile idiopathic arthritis (JIA). It is designed to capture beliefs related to disease management as well as social and emotional issues. The instrument was validated in children ages 7–17 years in 2001 (15); a Finnish translation was validated in 2007 with children ages 8–17 years (47).

Content. Items tap confidence in the ability to manage symptoms ("hurt," "tiredness"), emotions ("sad," "annoyed or fed-up"), and social participation ("at school," "with my friends"). Principal component factor analysis confirmed this 3-factor structure.

Number of items. 11 items total, in subscales: activity (4 items), symptom (4 items), and emotions (3 items).

Response options/scale. Five-point scale, from 1 ("not at all sure") to 5 ("very sure"). Item stem for each question is either "I can find ways to" control the hurt of arthritis, stop arthritis from making me feel sad; or "I can control my arthritis" "at school," "when I go out with my family." Finnish translation used not at all certain to very certain. Scale is 1–5 for each subscale.

Recall period for items. Not specified: mark which "describes you the best."

Endorsements. There are no identified endorsements.

Examples of use. CASE has been used in an evaluation of an internet-based self-management education program for adolescents (48).

Practical Application

How to obtain. Contact Julie Barlow, BA, PhD, Applied Research Centre in Health and Lifestyle Interventions, School of Health and Life Sciences, Coventry University, Priory Street, Coventry CV1 5FB, UK. Phone: 024 7688

7452. E-mail: j.barlow@coventry.ac.uk. Entire scale is published as appendix in original validation article (15).

Method of administration. Self-administered written self-report.

Scoring. Mean scores are calculated manually for each subscale using simple addition and division. Authors also calculated standard scores on a 0–10 scale to allow comparisons across subscales. No specific instructions are provided for handling missing values.

Score interpretation. Range for each subscale is 1–5 with higher scores indicating greater efficacy. No cut points or norms are available, but means are provided for each subscale in the original publication and the Finnish revalidation.

Respondent burden. Not specified, but estimated at <5 minutes. Items use language familiar to children and item length is short to ease readability. Topics are relevant to living with JIA and do not appear sensitive.

Administrative burden. Takes approximately 5 minutes to administer; scoring can be done manually by calculating means for each subscale. No training is required.

Translations/adaptations. Finnish translation was validated in 2007 (47).

Psychometric Information

Method of development. Items were developed based on literature review and focus groups with 5 subgroups: children with mild or severe JIA, parents of children with mild or severe JIA, and health professionals. Eleven issues emerged as salient to children with JIA. Items were written in language the children used. Subscales were developed using principal component factor analysis and explain 76.5% of the variation. Item response theory was not used in scale development.

Acceptability. The CASE was pilot tested for ease of use and comprehensibility before the validation study; no problems were noted. Readability appears appropriate for children; it is not clear if there are ceiling or floor effects, or whether missing data are common.

Reliability. All 3 subscales showed reasonable internal consistency reliability in the initial validation (Cronbach's alpha ranging from 0.85–0.90) and in the Finnish revalidation (Cronbach's alpha ranging from 0.77–0.80).

Validity. In terms of construct validity, CASE correlated significantly with theoretically relevant variables; positive correlations were found with hope and physical and psychological well-being, and negative correlations were found with measures of function, anxiety, pain, fatigue, and stiffness. In terms of criterion validity, CASE subscales had positive correlations with the Children's Hope scale, identified as a measure of control: activity subscale ($r = 0.56$), symptoms subscale ($r = 0.56$), and emotions subscale ($r = 0.61$). The Finnish revalidation study found all correlations in the expected direction as well (2). Finnish revalidation study also generally confirmed the factor structure of the CASE; in both validation studies, 1 item (swollen joints) loaded on both the symptoms and emotions subscales.

Ability to detect change. Unknown.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CASE is the only arthritis-specific measure of self-efficacy for children with arthritis, and self-efficacy is assumed to be an important factor in managing juvenile arthritis. Although the initial validation study was small (89 children from a single hospital specialty clinic in the UK), the revalidation of a Finnish translation used a slightly larger sample size (120 children from a rheumatology specialty hospital). In both studies, there could be bias toward children that need specialty care, but the Finnish investigators reported that the catchment area for the hospital was the entire country, suggesting a more population-based sample. The CASE has reasonably good internal consistency reliability and construct validity.

Caveats and cautions. Evidence on test–retest reliability and sensitivity to change would strengthen confidence in the use of this measure. The CASE developers raise concerns that the beliefs and expectations of adolescents with JIA may be different than younger children, but there is no adolescent arthritis self-efficacy measure in the literature.

Clinical usability. Both the respondent and administrative burdens are reasonable for clinical use. CASE developers stated that it was developed to help understand variations in adjustment to JIA and that it may assist in identifying children at risk for poor adjustment (based on low self-efficacy), but the lack of population norms will limit clinicians' ability to draw conclusions about individuals based on their CASE scores.

Research usability. Both the respondent and administrative burdens are reasonable for use in research. CASE developers stated that a second reason for development was to serve as an outcome measure in evaluation of psychoeducational interventions. Test–retest reliability and information on sensitivity to change could strengthen CASE for use in intervention research.

CHRONIC DISEASE SELF-EFFICACY SCALE (CDSES)

Description

Purpose. The CDSES was developed to assist in program evaluation of Stanford's Chronic Disease Self-Management Program. The development and testing of the CDSES is detailed in a book by Lorig et al published in 1996 (14); full psychometric data have not been published in the peer-reviewed literature. The authors describe self-efficacy as a belief in one's ability to use those skills in realistic contexts, and a belief that the use of the skills will produce the desired outcomes. The authors delineated 3 types of self-efficacy beliefs (to perform specific behaviors, to manage disease generally, and to achieve outcomes); a total of 10 subscales, each ranging from 1–10 items, is included in the original CDSES. More recently, a shortened version, labeled the Self-Efficacy for Managing Chronic Disease 6-item Scale, has been used. This 6-item scale combines items from 2 of the original 10 subscales (49,50).

Content. The full CDSES measures multiple diverse aspects of managing chronic disease (see description of subscales below). Some items are specific to a behavior (do aerobic exercise 3–4 times per week, ask your doctor about things that concern you), and some items assess confidence in attaining a result (get friends to help you, reduce emotional distress, keep fatigue from interfering with things you want to do). Items on the 6-item scale pertain primarily to performance accomplishment rather than behavior (keep various symptoms from interfering with things you want to do).

Number of items. The original CDSES, in its entirety, contains 33 items in 10 subscales. It is not clear if any studies have used all 10 subscales. The subscales are conceptually divided into 3 types of self-efficacy, as follows: self-efficacy to perform self-management behaviors (exercise regularly [3 items], get information about disease [1 item], obtain help from community, family, friends [4 items], communicate with physician [3 items]), self-efficacy to manage disease in general (manage disease in general [5 items]), and self-efficacy to achieve outcomes (do chores [3 items], social/recreational activities [2 items], manage symptoms [5 items], manage shortness of breath [1 item], control/manage depression [6 items]). The shortened version, Self-Efficacy for Managing Chronic Disease 6-item Scale, contains 3 items from the manage symptoms subscale and 3 from the manage disease in general subscale. No published psychometric data are available on this shortened version; mean, SD, and internal consistency reliability is reported for an undescribed sample of 605 people with chronic disease on the developer web site.

Response options/scale. Item stem for each item is "How confident are you that you can . . ." Responses are a 1–10 numerical rating scale for each item (1 = not at all confident, 10 = totally confident).

Recall period for items. "At the present time."

Endorsements. There are no known endorsements.

Examples of use. The CDSES, in its various lengths, has been used in the evaluation of self-management education program, primarily the Chronic Disease Self Management Program in its various forms (49–58). The full 33-item measure has been used (51,52), as well as select subscales of the full scale (53,54) and shortened versions (49,50,55).

Practical Application

How to obtain. The full scale and shortened scale are available from the web site of the Stanford Patient Education Research Center, URL: http://patienteducation.stanford.edu/.

Method of administration. Written questionnaire; self-administered.

Scoring. Scoring instructions are provided, including how to handle missing items. Computer scoring is not required; scoring involves addition and division to calculate mean score for each subscale used.

Score interpretation. Each subscale score range is 1–10. There are no population-based norms or cut-off scores, although scale documentation does provide mean scores and SDs for each subscale for respondents participating in the validation sample (ranging from 280–478 per subscale).

Respondent burden. Time to complete will depend on number of subscales used. The shortened 6-item version can be completed in approximately 3 minutes. Reading level is not difficult but some items are complex (i.e., "how confident are you that you can do things other than just take medications to reduce how much your illness affects your everyday life?").

Administrative burden. No training is required to administer the CDSES; scoring time will depend on the number of subscales used, but each subscale should take <5 minutes.

Translations/adaptations. A 4-item Spanish CDSES is available (55); the developers state that it was developed and tested in Spanish (rather than translated from English), but very limited psychometric data are available on the developer web site.

Psychometric Information

Method of development. Items were generated as a result of literature review that identified 12 self-management tasks common across chronic conditions, and 11 focus groups where participants were asked to describe their experiences and perceptions. Subcategories of self-efficacy (for self-management behaviors, manage disease in general, and to achieve outcomes) were delineated conceptually. It is not clear how the individual subscales were created, but the developer reports using multi-trait scaling approaches to assure correlation of an item to its designated subscale and limited correlation with other subscales.

Acceptability. Reading level seems appropriate, although some items appear complex. The authors report no ceiling or floor effects. It is not clear how much missing data occurred.

Reliability. Initial validation of the CDSES was drawn from a number of respondents in the Chronic Disease Self-Management Program intervention trials (total n = 1,130, although no subscale had more than 478 respondents). Internal consistency coefficients ranged from 0.77–0.92 among the various subscales of the CDSES for this accumulated convenience sample. Stability was measured in a small sample (51 respondents) in a 10-day test–retest procedure. Test–retest correlations were 0.82–0.89 for the different subscales. Internal consistency coefficient for the 6-item shortened scale was reported as 0.91 on the web site, but no publication is cited. Data are derived from an undescribed sample of 605 people with chronic disease. No test–retest coefficient is reported for the shortened scale. In the randomized controlled trial of the Spanish intervention program, the Spanish CDSES had an internal consistency alpha coeffiecient of 0.85 (n = 147) and test–retest coefficient of 0.80.

Validity. Limited data on validity are available. The publication includes a correlation matrix showing that the subscales of the full measure are correlated between 0.14 (self-efficacy to manage shortness of breath with do chores) and 0.68 (self-efficacy to manage symptoms with manage depression). The self-efficacy to manage disease subscale has correlated more strongly with the other subscales, which the developer expects since it is more of a summary measure. No data are reported to demonstrate correlations of the CDSES with other measures of self-efficacy to evaluate criterion validity. Modest support for construct validity can be derived from the correlations between self-efficacy subscales and their corresponding health behavior scales (ranging from 0.01–0.41) and correlations between health outcomes and self-efficacy subscales (range from 0.14–0.75). No intervariable correlations are available to examine the validity of the shortened 6-item scale or the Spanish version of the scale.

Ability to detect change. Sensitivity to change is not addressed in the CDSES documentation, although intervention studies do show changes in self-efficacy scores.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CDSES is widely used in evaluation of generic chronic disease self-management interventions, which often include people with arthritis. The 10 subscales of the full CDSES would provide the opportunity to assess a broad array of self-efficacy beliefs related to management of chronic disease.

Caveats and cautions. No psychometric data on the CDSES have been published in peer-reviewed publications. While the authors make conceptual distinctions among self-efficacy to perform specific behaviors from self-efficacy to manage the disease in general, and self-efficacy to achieve outcomes, some of the items in the behavior-specific subscales appear to tap concepts that could be considered outcomes of behavior rather than behavior itself (i.e., “get family and friends to help you”). It is also unclear how similar or different self-efficacy to achieve outcomes is from outcome expectations. There are limited validity data on the shortened 6-item scale, which is probably used most commonly, or the Spanish 4-item scale. Finally, the introduction of the shortened 6-item scale has also introduced some variability and confusion into the literature. Some investigators use the 6-item shortened scale (a composite of 3 items from the managing disease in general and 3 items from the managing symptoms subscales of the full scale) (49,50), while others use just the 5-item managing disease in general subscale (55); in both cases, it is usually called chronic disease self-efficacy. Other articles just refer to the CDSES developed at Stanford or for the Chronic Disease Self-Management Program without specifying the number of items, so it is not clear which iteration was used (56,57).

Clinical usability. Although there are means and SDs for scores from the accumulation of respondents used for the psychometric report, there are no published population-based norms or cut-off points, which makes it difficult to interpret individual scores. Respondent and administrative burden could be a problem for the full scale, but the shortened scales should not create a problem.

Research usability. Select subscales of the CDSES, or its 6-item shortened version, have been used extensively in research on the Chronic Disease Self-Management Program. Most investigators have used selected subscales or the shortened 6-item scale.

PARENT’S ARTHRITIS SELF-EFFICACY SCALE (PASE)

Description

Purpose. The PASE was designed to measure mothers’ and fathers’ perceived ability to manage or control salient aspects of their school-aged child’s juvenile idiopathic arthritis (JIA) (16). The psychometrics of the scale were reported separately for mothers and fathers. It is challenging to place the PASE in the context of self-efficacy measures because it asks individuals (parents) to estimate how certain they are that they can control aspects of another person’s (their child’s) arthritis, in contrast to usual self-efficacy measures that question a person’s confidence in their own ability to perform a specific action. The scale is not a proxy measure (e.g., asking parents to estimate their child’s efficacy) but is measuring the parent’s self-efficacy for an arthritis-specific parenting task. It is important to note that the scale is based on the hypothesis that a parent’s health status is influenced by their perceived ability to handle a specific parenting task, that is, managing their child’s arthritis. It was hypothesized, secondarily, that the parental sense of competence would influence the child’s physical and psychological health status, but self-efficacy theory does not seem the basis for this theoretical formulation. The measure was originally published in 2000, with a Finnish translation and revalidation published in 2007 (47).

Content. Items reflect 14 issues found to be salient in preliminary research. These include management of pain, stiffness, swelling, fatigue, sleep, loneliness, frustration, pleasure, and participation in school, family, and friend activities. Where content was similar, items were modifications of Arthritis Self-Efficacy Scale items. Item example: “How certain are you that you can keep arthritis pain from interfering with your child’s sleep?”

Number of items. 14 items total; initial validation study principal component factor analysis revealed 2 subscales (symptoms and psychosocial), each consisting of 7 items. In the validation of the Finnish translation, this 2-factor solution was not supported and a 3-factor model emerged (somatic [5 items], psychological [5 items], and social [4 items]). However, this analysis was done combining mothers’ and fathers’ responses, where the original factor analysis separated mothers and fathers.

Response options/scale. A 7-point response scale from 1 (very uncertain) to 7 (very certain), and a not applicable category, in response to items that begin “How certain are you that you can. . . .”

Recall period for items. “At the present time.”

Endorsements. There are no known endorsements identified.

Examples of use. There are no references located beyond the psychometric studies.

Practical Application

How to obtain. Contact Julie Barlow, BA, PhD, Applied Research Centre in Health and Lifestyle Interventions, School of Health and Life Sciences, Coventry University, Priory Street, Coventry CV1 5FB, UK. Phone: 024 7688 7452. E-mail: j.barlow@coventry.ac.uk. The entire scale is published as an appendix in the original validation article (16).

Method of administration. Written, self-administered self-report that is easy to administer.

Scoring. Sum of item scores on each subscale; this can be done manually. Validation study also standardized scores to a 0–10 scale allowing easier comparison across subscales. This would be labor intensive if attempted manually. There are no instructions for handling missing values or "not applicable" responses.

Score interpretation. Total score range would be 7–49 for each subscale, assuming no "not applicable" responses were recorded. Higher scores reflect greater confidence in ability to manage or control aspects of child's juvenile arthritis. No cut points or population norms are provided, although the initial validation and Finnish translation revalidation do report mean scores and SDs for each subscale for both mothers and fathers.

Respondent burden. Not reported; estimated to be <5 minutes. Items appear easy to read and not sensitive.

Administrative burden. Administration time is likely to be rapid; manual scoring or mean subscale scores should be rapid, although calculation of standard scores could be more time consuming. No training is required.

Translations/adaptations. A Finnish translation was validated in 2007 (47).

Psychometric Information

Method of development. Items were generated by instrument developers based on past experience, literature review on self-efficacy and impact of arthritis on parenting, and focus groups with 5 subpopulations (children with mild or severe JIA, parents of children with mild or severe JIA, and health professionals). Principal component factor analysis was used to generate the subscales, which explain 75.5% of variance for mothers and 65.8% of variance for fathers.

Acceptability. The PASE was pilot tested for ease of use and comprehensibility by parents of 13 children with JIA; no problems were noted. It is not known if there are ceiling or floor effects, or if missing data are common.

Reliability. Questionnaires were sent to 149 families from 2 hospitals in the UK. A total of 178 parents participated in the validation study. In the initial validation study, internal consistency reliability was reasonable, with Cronbach's alpha for mothers ranging between 0.92 (symptoms subscale) and 0.96 (psychological subscale); Cronbach's alpha for fathers ranged between 0.89 (symptoms subscale) and 0.93 (psychological subscale). In the Finnish translation and revalidation study, internal consistency reliability was conducted for mothers and fathers combined for the 3 subscales that emerged; Cronbach's alpha ranged from 0.84 (somatic) to 0.88 (psychological) and 0.93 (social). No test–retest reliability data are available.

Validity. Criterion validity was demonstrated by significant correlations with the Generalized Self-Efficacy Scale (a general measure of perceived coping competence) with both subscales of the PASE, for both mothers and fathers (0.27 for symptoms subscale for mothers, 0.36 for symptoms subscale for fathers, 0.43 for psychosocial subscale for mothers, and 0.33 for psychosocial subscale for fathers). Construct validity was demonstrated for mothers by significant negative association of mothers' anxious and depressed mood with symptoms subscale ($r = -0.28$) and psychosocial subscale ($r = -0.43$), and significant associations of mothers' psychosocial efficacy with their physical function ($r = 0.27$), energy ($r = 0.27$), pain ($r = 0.31$), and general health perceptions ($r = 0.37$). The only significant associations for fathers were positive associations between fathers' general health perceptions and psychosocial subscale ($r = 0.31$), and negative association between fathers' depressed mood and psychosocial subscale ($r = -0.29$). Authors also investigated the associations between parents' or child's ratings of child's physical and psychosocial well-being and parental self-efficacy ratings. Investigators specify that they expected parental self-efficacy to be reflected in child's well-being, but did not provide strong theoretical rationale for including this as evidence of construct validity.

Ability to detect change. No evidence of ability to detect change is available.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PASE purports to measure a factor (parents' perceived competence in managing their child's JIA) that may be related to the physical and psychological health statuses of children with JIA and their parents. This could help increase understanding of family adaptation to JIA by measuring parental self-efficacy. This could also be useful in considering the family as a unit, rather than just considering the child in isolation. Internal consistency reliability for both subscales was strong, and the measure performs in theoretically consistent ways for mothers but not fathers. The sample for the initial validation study was 178 parents (115 mothers, 63 fathers) drawn from 2 hospitals in the UK; however, the authors state those hospitals draw from a wide geographic area, which could widen its generalizability.

Caveats and cautions. There is no information about test–retest reliability or sensitivity to change which can limit its usefulness in evaluating interventions. Similar to the Arthritis Self-Efficacy Scale on which it was modeled, the PASE focuses on performance attainments (decrease child's pain), which some might consider an outcome expectation. Since its initial publication in 2000, it does not appear to have been used in any published studies beyond the Finnish translation and re-validation, which makes it unclear if investigators will find it useful. It is noteworthy that the original validation study identified a 2-factor structure, while the Finnish revalidation study identified a 3-factor structure.

Clinical usability. Both the respondent and administrative burdens are reasonable for clinical use. PASE developers stated that it may be useful to identify mothers at risk of poor adjustment due to their child's JIA, and may help understand variations in family adjustment, but the lack of population norms will limit clinicians' ability to draw conclusions about mothers or families based on their PASE scores.

Research usability. Both the respondent and administrative burdens are reasonable for use in research. The PASE could be useful for evaluating parent-oriented interventions to improve management of or coping with JIA, but information on sensitivity to change and test–retest reliability would be necessary.

RHEUMATOID ARTHRITIS SELF-EFFICACY SCALE (RASE)

Description

Purpose. The RASE was developed to measure task-specific self-efficacy for the initiation of self-management–related behavior, and was carefully worded to tap beliefs about capability to perform the behavior, rather than actual ability, performance, or outcome expectation. It was developed specifically for patients with rheumatoid arthritis (RA) in the UK, although the items reflect self-management behaviors important in self-management of other forms of arthritis as well. The original validation study was published in 2001 (13), with a revalidation published in 2008 (9), and a Danish translation published in 2010 (58). A shortened version of the RASE based on an item response theory (IRT) analysis in the US has also been proposed (38).

Content. Self-efficacy for self-management in RA is assumed to be a multidimensional concept; items address beliefs about ability to perform tasks across 8 dimensions of self-management identified as important in RA (relaxation, relationships, function, leisure activities, exercise, sleep, medication, and fatigue). However, these dimensions are all summed into a single factor, rather than subscales. The authors note that some people will have high self-efficacy for some tasks and low self-efficacy for others.

Number of items. 28. No subscales are used, although factor analysis showed 8 factors explaining 75% of the variance. A shortened RASE (9 items) has been proposed based on a content- and statistics-driven IRT analysis, which produced a scale of modest reliability (0.84) (38).

Response options/scale. Item stem is "I believe I *could*" with response options 1 (strongly disagree) to 5 (strongly agree).

Recall period for items. Not specified.

Endorsements. There are no noted endorsements.

Examples of use. RASE has been used in evaluations of an arthritis education program (59) and a physical activity program (60).

Practical Application

How to obtain. Contact Sarah Hewlett, PhD, MA, RN, Professor of Rheumatology and Nursing, Academic Rheumatology, Bristol Royal Infirmary, Bristol BS2 8HW, UK. Phone: 44 (0) 117 928 2903. Fax: 44 (0) 117 928 3841. E-mail: Sarah.Hewlett@uwe.ac.uk. The full instrument is published as Appendix 1 in the original validation study and 2008 revalidation (9,13).

Method of administration. Self-administered written self-report questionnaire.

Scoring. Scoring is simple addition of responses; no computer is necessary. There are no instructions for handling missing items.

Score interpretation. Score range is 28–140 with higher scores indicating greater self-efficacy. No cut-off scores or population norms are available.

Respondent burden. Not reported; estimated to be <10 minutes.

Administrative burden. No training is required to administer the RASE; scoring time is the time to add 28 items.

Translations/adaptations. A Danish translation was published in 2010 (58).

Psychometric Information

Method of development. A multi-stage process was used for item generation. Initial items emerged from interviews with 19 health professionals and 17 people with RA. The original pool of 166 items was reduced to 100 items by examination of frequency of mention and designation of helpfulness by the 17 people with arthritis. The 100-item initial questionnaire was pilot tested with 92 people with RA in the UK. Three sets of analyses (item correlation with other SE items or clinical and psychological variables, principal component analysis, and correlations of each item with mean RASE score) were used to pare the original 100 items to the 28 items included in the final RASE. A separate IRT analysis was conducted by Mielenz et al using data gathered from educated white women in the US with a variety of types of arthritis. This analysis proposed a 9-item shortened RASE that they report is representative of the construct of interest, which they describe as a common construct with 9 subfactors. Item selection for this shortened version relied on both content and statistical analysis, with IRT analysis used to support this conceptualization (38).

Acceptability. Items do not appear difficult to complete; in the initial validation, 85% of the respondents completed more than 90% of the items and no item was consistently omitted.

Reliability. Initial reliability and validity testing utilized outpatients with RA (n = 107). Cronbach's alpha showed good internal consistency reliability in the 2008 revalidation (0.89) and moderate to strong correlation of each item to the total RASE score. Four-week test–retest reliability was also good (0.90 in the initial validation study). Confirmatory factor analysis in the 2008 revalidation (n = 128 people with RA who enrolled in an education program, from 11 treatment centers) showed similar factor loadings on the 8 factors identified in the initial validation study.

Validity. In terms of construct validity, as predicted by self-efficacy theory, the RASE is correlated with initiation of corresponding self-management behaviors (mean change score 5.4 points on 0–28 scale) following self-management education intervention. In terms of convergent validity, a significant correlation was found between the Arthritis Self-Efficacy Scale (ASES) other symptoms subscale and the RASE (r = 0.313); changes in the RASE were correlated with changes in the ASES pain subscale (r = 0.35) and ASES other symptoms subscale (r = 0.32). In terms of divergent validity, neither the RASE nor ASES showed significant correlation with the General Self-Efficacy Scale, a trait measure of optimistic self-beliefs and perceived coping competence (in contrast to the more behavior-specific concepts of the RASE and ASES).

Ability to detect change. RASE showed small but significant changes in SE following participation in a variety of self-management education programs in the UK (mean change 5.2 points on scale scoring 28–140). The standardized response means showed sensitivity to change whether calculated as absolute change or percentage change, and were similar in both the 2-week and 8-week post intervention analyses.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The RASE is a measure of self-efficacy beliefs related to self-management behaviors in RA. Although it was developed specifically for RA patients in the UK, the items appear to be appropriate for other forms of arthritis, and other countries as well. Extensive psychometric evaluation has been conducted, and the RASE has good reliability, validity, and sensitivity to change. In contrast to the ASES, which includes items addressing specific functions ("walk 100 feet on flat ground in 20 seconds") and performance results ("decrease your pain quite a bit"), the RASE asks about ability to perform specific self-management behaviors ("use relaxation techniques to help with pain"). A shortened version of the RASE has been proposed following an IRT analysis, but no published use of this shortened version was located.

Caveats and cautions. The 28-item RASE has been validated exclusively in the UK; although it has been used in the US, there has been no psychometric analysis to confirm its appropriateness. Similarly, the majority of the psychometric analysis has used in people with RA; although the RASE has been used in community samples of people with arthritis, the instrument has not been revalidated with this more broad population.

Clinical usability. Neither the administrative nor respondent burden should preclude its use, but the absence of any population-based norms or cut-off scores makes it difficult to interpret an individual's score.

Research usability. The available psychometric data, including good reliability, validity, and reasonable sensitivity to change, all suggest the RASE is appropriate for use in research. Neither the administrative nor respondent burden should preclude its use.

AUTHOR CONTRIBUTIONS

Dr. Brady drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Allegrante JP, Marks R. Self-efficacy in management of osteoarthritis. Rheum Dis Clin North Am 2003;29:747–68.
2. Lefebvre JC, Keefe FJ, Affleck G, Raezer LB, Starr K, Caldwell DS, et al. The relationship of arthritis self-efficacy to daily pain, daily mood, and daily pain coping in rheumatoid arthritis patients. Pain 1999;80:425–35.
3. Cross MJ, March LM, Lapsley HM, Byrne E, Brooks PM. Patient self-efficacy and health locus of control: relationships with health status and arthritis-related expenditure. Rheumatology (Oxford) 2006;45:92–6.
4. Brekke M, Hjortdahl P, Kvien TK. Self-efficacy and health status in rheumatoid arthritis: a two-year longitudinal study. Rheumatology (Oxford) 2001;40:387–92.
5. Brus H, van de Laar M, Taal E, Rasker J, Wiegman O. Determinants of compliance with medication in patients with rheumatoid arthritis: the importance of self-efficacy expectations. Patient Educ Couns 1999;36:57–64.
6. Taal E, Rasker JJ, Seydel ER, Wiegman O. Health status, adherence with health recommendations, self-efficacy and social support in patients with rheumatoid arthritis. Patient Educ Couns 1993;20:63–76.
7. Bandura A. Self-efficacy: toward a unifying theory of behavior change. Psychol Rev 1977;84:191–215.
8. Bandura A. Self-efficacy: the exercise of control. New York: W.H. Freeman; 1997.
9. Hewlett S, Cockshott Z, Almeida C, Richards P, Lowe R, Greenwood R, et al. Sensitivity to change of the RA self-efficacy scale (RASE) and predictors of change in self-efficacy. Musculoskeletal Care 2008;6:49–67.
10. Gecht MR, Connell KJ, Sinacore JM, Prohaska TR. A survey of exercise beliefs and exercise habits among people with arthritis. Arthritis Care Res 1996;9:82–8.
11. Frei A, Svarin A, Steurer-Stey C, Puhan MO. Self-efficacy instruments for patients with chronic diseases suffer from methodological limitations: a systematic review. Health Qual Life Outcomes 2009;7:86.
12. Lorig K, Chastain RL, Ung E, Shoor S, Holman HR. Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. Arthritis Rheum 1989;32:37–44.
13. Hewlett S, Cockshott Z, Kirwan J, Barrett J, Stamp J, Haslock I. Development and validation of a self-efficacy scale for use in British patients with rheumatoid arthritis. Rheumatology (Oxford) 2001;40:1221–30.
14. Lorig K, Stewart A, Ritter P, Gonzalez V, Laurent D, Lynch J. Outcome measures for health education and other health care interventions. Thousand Oaks (CA): Sage Publications; 1996.
15. Barlow JH, Shaw KL, Wright CC. Development and preliminary validation of a children's arthritis self-efficacy scale. Arthritis Rheum 2001;45:159–66.
16. Barlow JH, Shaw KL, Wright CC: Development and preliminary validation of a self-efficacy measure for use among parents of children with juvenile idiopathic arthritis. Arthritis Care Res 2000;13:227–36.
17. McAuley E, Courneya KS, Lettunich J. Effects of acute and long term exercise on self-efficacy responses in sedentary middle-aged adults. Gerontologist 1991;31:534–42.
18. Bandura A. Guide for constructing self-efficacy scales. In: Urdan T, Parjares F, editors. Self-efficacy beliefs in adolescents. Charlotte: Information Age Publishing; 2006.
19. Thomee P, Waahrborg P, Boorjesson M, Thomee R, Eriksson BI, Karlsson J. A new instrument for measuring self-efficacy in patients with an anterior cruciate ligament injury. Scand J Med Sci Sports 2006;16:181–7.
20. Anderson KO, Dowds BN, Pelletz RE, Edwards WT, Peeters-Asdourian C. Development and initial validation of a scale to measure self-efficacy beliefs in persons with chronic pain. Pain 1995;63:77–84.
21. Levin JB, Lofland KR, Cassisi JE, Poreh AM, Blonsky ER. The relationship between self-efficacy and disability in chronic low back pain. Int J Rehabil Health 1996;2:19–28.
22. Altmaier EM, Russell DW, Feng Kao C, Lehmann TR, Weinstein JN. Role of self-efficacy in rehabilitation outcome among chronic low back pain patients. J Couns Psychol 1993;40:335–9.
23. Van Hartingsveld F, Ostelo RW, Cuijpers P, de Vos R, Riphagen II, de Vet HC. Treatment-related and patient-related expectations of patients with musculoskeletal disorders: a systematic review of published measurement tools. Clin J Pain 2010;26:470–88.
24. Barlow JH, Williams B, Wright C. The generalized self-efficacy scale in people with arthritis. Arthritis Care Res 1996;9:189–96.
25. Scholz U, Dona BG, Sud S, Schwarzer R. Is general self-efficacy a universal construct? Eur J Psychol Assess 2002;18:242–51.
26. Sherer M, Maddux JE, Mercandante B, Prentice-Dunn S, Jacobs B, Rogers RW. The self-efficacy scale: construction and validation. Psychol Rep 1982,51:663–71.
27. Pearlin LI, Schooler I. The structure of coping. J Health Soc Behav 1978;19:2–21.

28. Barlow JH, Williams B, Wright CC. The reliability and validity of the arthritis self-efficacy scale in UK context. Psychol Health Med 1997;2:3–17.
29. Lomi C. Evaluation of a Swedish version of the Arthritis Self-efficacy Scale. Scand J Caring Sci 1992;6:131–8.
30. Taal E, Riemsma RP, Brus HL, Seydel ER, Rasker JJ, Wiegman O. Group education for patients with rheumatoid arthritis. Patient Educ Couns 1993;20:177–87.
31. Unsal A, Kasikci MK. Effect of education on perceived self-efficacy for individuals with arthritis. Int J Caring Sci 2010;3:3–11.
32. Barlow JH, Turner AP, Wright CC. A randomized controlled study of the Arthritis Self-Management Programme in the UK. Health Educ Res 2000;15:665–80.
33. Buszewicz M, Rait G, Griffin M, Nazareth I, Patel A, Atkinson A, et al. Self management of arthritis in primary care: randomised controlled trial. BMJ 2006;333:879.
34. McKnight PE, Afram A, Kashdan TB, Kasle S, Zautra A. Coping self-efficacy as a mediator between catastrophizing and physical functioning: treatment target selection in an osteoarthritis sample. J Behav Med 2010;33:329–49.
35. Riemsma RP, Rasker JJ, Taal E, Griep EN, Wouters JM, Wiegman O. Fatigue in rheumatoid arthritis: the role of self-efficacy and problematic social support. Br J Rheumatol 1998;37:1042–6.
36. Shelby RA, Somers TJ, Keefe FJ, Pells JJ, Dixon KE, Blumenthal JA. Domain specific self-efficacy mediates the impact of pain catastrophizing on pain and disability in overweight and obese osteoarthritis patients. J Pain 2008;9:912–9.
37. Lomi C, Nordholm LA. Validation of a Swedish version of the arthritis self-efficacy scale. Scand J Rheumatol 1992;21:231–7.
38. Mielenz TJ, Edwards MC, Callahan LC. Item response theory analysis of two questionnaire measures of arthritis-related self-efficacy beliefs from community-based US samples. Arthritis 2010:416796.
39. Gonzales VM, Steward A, Ritter PL, Lorig K. Translation and validation of arthritis outcome measures into Spanish. Arthritis Rheum 1995;38:1429–46.
40. Mueller A, Hartmann M, Mueller K, Eich W. Validation of the arthritis self-efficacy short form scale in German fibromyalgia patients. Eur J Pain 2003;7:163–71.
41. Goeppinger J, Lorig KR, Ritter PL, Mutatkar S, Villa F, Gizlice Z. Mail-delivered arthritis self-management tool kit: a randomized trial and longitudinal followup. Arthritis Rheum 2009;61:867–75.
42. Lorig K, Gonzalez VM, Ritter P. Community-based Spanish language arthritis education program: a randomized trial. Med Care 1999;37:957–63.
43. Osborne R, Wilson T, Lorig K, McColl G. Does self-management lead to sustainable health benefits in people with arthritis? A 2-year transition study of 452 Australians. J Rheumatol 2007;34:1112–7.
44. Cadmus L, Patrick MB, Maciejewski ML, Topolski T, Belza B, Patrick DL. Community-based aquatic exercise and quality of life in persons with osteoarthritis. Med Sci Sports Exerc 2010;42:8–15.
45. Turner JA, Ersek M, Kemp C. Self-efficacy for managing pain is associated with disability, depression, and pain coping among retirement community residents with chronic pain. J Pain 2005;7:471–9.
46. Allen KD, Oddone EZ, Coffman CJ, Keefe FJ, Lindquist JH, Bosworth HB. Racial differences in osteoarthritis pain and function: potential explanatory factors. Osteoarthritis Cartilage 2010;18:160–7.
47. Vuorimaa H, Honkanen V, Konttinen YT, Komulainen E, Santavirta N. Improved factor structure for self-efficacy scales for children with JIA (CASE) and their parents (PASE). Clin Exp Rheumatol 2007;25:494–501.
48. Stinson JN, McGrath PJ, Hodnett ED, Feldman BM, Duffy CM, Huber AM, et al. An internet-based self management program with telephone support for adolescents with arthritis: a pilot randomized controlled trial. J Rheumatol 2010;37:1944–52.
49. Gitlin L, Chernett N, Harris LF, Palmer D, Hopkins P, Dennis MP. Harvest health: translation of the chronic disease self management program for older African Americans in a senior setting. Gerontologist 2008;48:698–705.
50. Griffiths C, Motlib J, Azad A, Ramsay J, Eldridge S, Feder G. Randomized controlled trial of a lay-led self management programme for Bangladeshi patients with chronic disease. Br J Gen Pract 2005;55:831–7.
51. Jerant A, Moore M, Lorig K, Franks P. Perceived control moderated the self efficacy-enhancing effects of a chronic illness self-management intervention. Chronic Illn 2008;4:173–82.
52. Siu AM, Chan CC, Poon PK, Chui DY, Chan SC. Evaluation of the chronic disease self management program in a chinese population. Patient Educ Couns 2007;65:42–50.
53. Kennedy A, Reeves D, Bower P, Lee V, Middleton E, Richardson G, et al. The effectiveness and cost effectiveness of a national lay-led self care support programme for patients with long-term conditions: a pragmatic randomised controlled trial. J Epidemiol Community Health 2007;61:254–61.
54. Barlow JH, Wright CC, Turner AP, Bancroft GV. A 12-month follow-up study of self-management training for people with chronic disease: are changes maintained over time? Br J Health Psychol 2005;10:589–99.
55. Lorig KR, Ritter PL, Gonzalez VM. Hispanic chronic disease self-management: a randomized community-based outcome trial. Nurs Res 2003;52:361–9.
56. Swerissen H, Belfrage J, Weeks A, Jordan L, Walker C, Furler J, et al. A randomised control trial of a self-management program for people with a chronic illness from Vietnamese, Chinese, Italian and Greek backgrounds. Patient Educ Couns 2006;64:360–8.
57. Lorig KR, Ritter PL, Dost A, Plant K, Laurent DD, McNiel I. The expert patients programme online: a 1-year study of an internet-based self-management programme for people with long-term conditions. Chronic Illn 2008;4:247–56.
58. Primdahl J, Wagner L, Horslev-Peterson K. Self-efficacy in rheumatoid arthritis: translation and test of validity, reliability, and sensitivity of the Danish version of the rheumatoid arthritis self-efficacy questionnaire (RASE). Musculoskeletal Care 2010;8:123–35.
59. Hammond A, Bryan J, Hardy A. Effects of a modular behavioural arthritis education programme: a pragmatic parallel-group randomized controlled trial. Rheumatology (Oxford) 2008;47:1712–8.
60. Callahan LF, Mielenz T, Freburger J, Shreffler J, Hootman J, Brady T, et al. A randomized controlled trial of the People with Arthritis Can Exercise program: symptoms, function, physical activity, and psychosocial outcomes. Arthritis Rheum 2008;59:92–101.

Summary Table for Self-Efficacy Measures

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
Arthritis Self-Efficacy Scale	Arthritis-specific measure of self-efficacy beliefs; originally designed to help explain changes resulting from arthritis education programs	Self-administered, self-report	Not reported, estimated to be 5–10 minutes	Time to administer and score 5–15 minutes; can be scored by hand	Range 10–100 or 1–10 for each subscale; higher scores indicate higher self-efficacy	Cronbach's alpha ranged 0.76–0.87 per subscale, test–retest ranged 0.85–0.90 per subscale	Construct validity: correlated in predicted ways with health status measures	Unknown, but changes reported in intervention studies	Demonstrates reasonable reliability and validity data and has demonstrated sensitivity to change in intervention studies; appropriate for use in intervention evaluation studies	Unclear if scales correlate with corresponding health behaviors; concerns about possible overlap between self-efficacy for physical function and actual function; no ability to interpret individual scores
Arthritis Self-Efficacy Scale-8 Item	Shortened version of the Arthritis Self-Efficacy Scale	Self-administered, self-report; original Spanish version is interviewer administered	Not reported, estimated to be <5 minutes	Time to administer and score 5–10 minutes; can be scored by hand	Range 1–10; higher scores indicate higher efficacy	Spanish version: Cronbach's alpha 0.92, test–retest 0.69; German translation: Cronbach's alpha 0.90, test–retest 0.51	No information on Spanish or English versions; construct validity: German version correlated with theoretically relevant variables	Not reported for Spanish or English versions, but changes reported in interventions studies; German version reported medium size changes	Shortened version is intuitively appealing for speed and ease of use; German translation showed good reliability and validity, and sensitivity to change	The limited psychometric data on the English translation are unpublished; no ability to interpret individual scores
Children's Arthritis Self-Efficacy Scale	Arthritis-specific measure designed to measure children's perceived ability to control or manage salient aspects of life with juvenile idiopathic arthritis	Self-administered, self-report	Not reported, estimated to be <5 minutes	Time to administer and score 5–10 minutes; can be scored by hand; creation of standard scores likely takes more time	Range 1–5 for each subscale, higher scores indicate higher efficacy	Cronbach's alpha ranged 0.85–0.90; Finnish translation Cronbach's alpha ranged 0.77–0.80	Construct validity: correlated with theoretically relevant health status variables	Unknown	Only arthritis-specific measure of self-efficacy for children; reasonably good reliability and validity demonstrated; could be used for evaluation of interventions	No test–retest reliability data; no ability to interpret individual scores
Chronic Disease Self-Efficacy Scales	Non–arthritis-specific measure of perceived ability to manage a chronic disease, to perform specific behaviors, and to achieve outcomes related to chronic disease management	Self-administered, self-report	Depends on number of subscales used, shortened 6-item version estimated to be 3 minutes	Time to administer and score depends on number of subscales used, 6-item version administered and scored in 5–7 minutes; can be scored by hand	Range 1–10 for each subscale; higher scores indicate higher self-efficacy	Cronbach's alphas ranged 0.77–0.92 for subscales, test–retest ranged 0.82–0.89; Cronbach's alpha for 6-item version 0.91, no test–retest reported	Construct validity: correlations among subscales and health status and health behaviors; no data on 6-item version	Unknown; changes reported in intervention studies	The menu of 10 subscales offers the opportunity to select most relevant subscales in research	Psychometric data for full scale or 6-item shortened scale have not been published in peer-reviewed publications; managing disease in general subscale can be confused with the 6-item measure; creation of the subscales is not defined; no ability to interpret individual scores
Parent's Arthritis Self-Efficacy Scale	Designed to assess arthritis-specific parenting challenges and perceived ability to manage or control their school-aged child's juvenile idiopathic arthritis	Self-administered, self-report	Not reported, estimated to be <5 minutes	Time to administer and score estimated to be 5-10 minutes; can be scored by hand	Range 7–49 for each subscale, higher scores indicate higher self-efficacy	Cronbach's alpha 0.92 and 0.96 for mothers; 0.89 and 0.93 for fathers; Finnish translation combined mothers and fathers for Cronbach's alpha 0.84–0.93	Construct validity: for mothers, correlations with select health status variables; for fathers, 1 significant correlation	Unknown	Only measure of parental efficacy in handling child's arthritis; good internal consistency reliability	No information of test–retest reliability or sensitivity to change; no ability to interpret individual scores
Rheumatoid Arthritis Self-Efficacy Scale	Rheumatoid arthritis–specific measure designed to measure initiation of arthritis self-management behaviors	Self-administered, self-report	Not reported, estimated to be <10 minutes	Time to administer and score estimated to be 10–15 minutes; can be scored by hand	Range 28–140; higher scores indicates greater efficacy	Cronbach's alpha 0.89; test–retest 0.90	Construct validity: correlations with initiation of health behaviors; convergent validity: modest correlations with subscales of Arthritis Self-Efficacy Scale	Detected change in intervention studies	Items focus on behavior-special conference; strong reliability, validity, and sensitivity to change	Validation studies conducted in rheumatoid arthritis; no ability to interpret individual scores

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S383–S412
DOI 10.1002/acr.20541
© 2011, American College of Rheumatology

Adult Measures of General Health and Health-Related Quality of Life

Medical Outcomes Study Short Form 36-Item (SF-36) and Short Form 12-Item (SF-12) Health Surveys, Nottingham Health Profile (NHP), Sickness Impact Profile (SIP), Medical Outcomes Study Short Form 6D (SF-6D), Health Utilities Index Mark 3 (HUI3), Quality of Well-Being Scale (QWB), and Assessment of Quality of Life (AQoL)

LUCY BUSIJA,[1] EVA PAUSENBERGER,[2] TERRY P. HAINES,[2] SHARON HAYMES,[1]
RACHELLE BUCHBINDER,[3] AND RICHARD H. OSBORNE[4]

INTRODUCTION

The aim of this review is to provide a summary of adult measures of general health and health-related quality of life (HRQOL) commonly used in rheumatology research studies. Currently, there is no single generally agreed upon definition or conceptual model of health or HRQOL, and developing a comprehensive definition of these complex concepts was beyond the scope of the review. For the purposes of this review, we define measures of general health and HRQOL as multi-item questionnaires that assess perceived health status and overall physical and emotional well-being that is not specific to any disease. The health measures included in this review were further subdivided into generic health profiles (questionnaires that provide assessment of more than 1 dimension of health status) and health utility measures that provide an overall measure of health status rated between perfect health (1.0) and death (0.0).

Relevant measures were identified (using medical subject headings [MeSH]) through a systematic search of medical publications indexed to PubMed database. The following search queries were used: [Quality of life (title) AND Outcomes assessment (MeSH terms) AND Rheumatic diseases (MeSH terms)] and [Quality of life (abstract) AND Outcomes assessment (MeSH terms) AND Rheumatic diseases (MeSH terms)]. In MeSH, “rheumatic diseases” are defined as “disorders of connective tissue, especially the joints and related structures, characterized by inflammation, degeneration, or metabolic derangement,” and include rheumatoid arthritis, Caplan’s syndrome, Sjögren’s syndrome, Still’s disease, fibromyalgia, gout, hyperostosis, osteoarthritis, and polymyalgia rheumatica among others.

Inclusion criteria were 1) the study was concerned with a rheumatology condition and 2) participants were human adults. The first query returned 129 items and the second query returned 494 items, with 623 abstracts in total. After removal of 77 duplicates, 38 pediatric studies, and 2 animal studies, 2 reviewers (LB, EP) screened abstracts independently of the remaining 506 publications to identify relevant multi-item questionnaires (i.e., those generic questionnaires that were identified by the study authors as being used for the purpose of assessing general health or HRQOL). Where abstracts contained insufficient information to determine the type of measures used, full publications were obtained.

The reviewers, working independently, identified 10 generic health utility measures and 5 generic health profiles (Table 1). Agreement about the type and number of occurrences of relevant measures in the sample of reviewed

Drs. Busija and Haymes’ work was supported by internal funding from the Centre for Eye Research Australia. Ms Pausenberger’s work was supported by internal funding from the Southern Health and Monash University Allied Health Research Unit. Dr. Haines’s work was supported by a National Health and Medical Research Council Career Development Award. Dr. Buchbinder’s work was supported in part by a National Health and Medical Research Council Practitioner Fellowship. Dr. Osborne’s work was supported in part by a National Health and Medical Research Council Population Health Career Development Award.

[1]Lucy Busija, PhD, Sharon Haymes, PhD: University of Melbourne, Melbourne, Victoria, Australia; [2]Eva Pausenberger, BPhysiother (Hons), Terry P. Haines, PhD: Southern Health, Cheltenham, Victoria, Australia, and Monash University, Melbourne, Victoria, Australia; [3]Rachelle Buchbinder, MBBS (Hons), FRACP, PhD: Cabrini Institute and Monash University, Melbourne, Victoria, Australia; [4]Richard H. Osborne, PhD: Deakin University, Melbourne, Victoria, Australia.

Dr. Buchbinder has received speaking fees (less than $10,000 each) from Pfizer, Wyeth, Abbott, and Roche, and (more than $10,000) from Abbott for participating as an international mentor for 3e Initiative.

Address correspondence to Lucy Busija, PhD, Center for Eye Research Australia, Royal Victorian Eye and Ear Hospital, 32 Gisborne Street, East Melbourne, VIC 3002, Australia. E-mail: lbusija@unimelb.edu.au.

Submitted for publication February 22, 2011; accepted in revised form June 20, 2011.

Table 1. Measures used for the assessment of general health and health-related quality of life in rheumatology literature

Questionnaire	Occurrences in reviewed abstracts, no. (n = 506)
Generic health profiles	
Short Form 36*	146
Nottingham Health Profile*	21
Short Form 12*	13
Sickness Impact Profile*	7
Duke Health Profile	1
Generic health utility measures	
EuroQol 5-domain†	32
Short Form 6D*	6
Health Utilities Index 3*	6
Quality of Wellbeing Scale (self-administered)*	5
Assessment of Quality of Life Scale*	4
15D	3
Quality of Life Scale	2
World Health Organization Quality of Life/Bref	2
Perceived Quality of Life Scale	1
Profile of Quality of Life in the Chronically Ill	1

* Review included in this article.
† Review of this measure is included in Measures of Disability article in this issue.

abstracts was very high (intraclass correlation coefficient 0.996). Any disagreements were resolved through a discussion between all authors. Given the large number of potentially relevant measures identified, only those measures that were used at least 4 times in the screened abstracts were selected for this review. Consequently, this report provides reviews of 4 generic health profiles: the Medical Outcomes Study Short Form 36 and Short Form 12 Health Surveys, the Nottingham Health Profile, and the Sickness Impact Profile, and 4 health utility measures: the Medical Outcomes Study Short Form 6D, the Health Utilities Index Mark 3, Quality of Well-Being Scale (self-administered), and the Assessment of Quality of Life Scale. Although the EuroQol 5-domain is also frequently used in rheumatology, the review of this measure is included in Measures of Disability article in this issue.

MEDICAL OUTCOMES STUDY SHORT FORM 36-ITEM (SF-36) AND SHORT FORM 12-ITEM (SF-12) HEALTH SURVEYS

Description

Purpose. The SF-36 and SF-12 are multi-item generic health surveys intended to measure "general health concepts not specific to any age, disease, or treatment group" (1). The SF-12 is a shorter version of the SF-36 and uses only 12 questions to measure functional health and well-being from the patient's perspective. The original objective was to develop a short, generic health-status measure that reproduces the 2 summary scores of the SF-36, i.e., the physical component summary (PCS) score and the mental component summary (MCS) score (2).

The SF-36 and SF-12 are suitable for use in general, as well as in clinical populations and, as such, can be used to compare health between populations and between diseases. The SF-36 and the SF-12 health surveys are available in original and revised versions. The SF-36 and SF-12 were first published in 1992 and 1996, respectively, with the revised versions of both questionnaires published in 2000. The revised versions are very similar to their original forms, with major differences involving changes in item wording, revision of the response scale to incorporate a greater number of response options, and norm-based scoring (3).

Content. Both the SF-36 and SF-12 measure 8 health domains: physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotional, and mental health. Physical functioning covers limitations in daily life due to health problems. The role physical scale measures role limitations due to physical health problems. The bodily pain scale assesses pain frequency and pain interference with usual roles. The general health scale measures individual perceptions of general health. The vitality scale assesses energy levels and fatigue. The social functioning scale measures the extent to which ill health interferes with social activities. The role emotional scale assesses role limitations due to emotional problems, and the mental health scale measures psychological distress.

The SF-36 and SF-12 can also be used to derive 2 aggregate summary measures: the PCS and the MCS. Summary scores are calculated by summing factor-weighted scores across all 8 subscales, with factor weights derived from a US-based general population sample (4). Country-specific weights are also available for Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, the UK (5), and Australia (6). In the calculation of the PCS summary score, highest weights are given to the physical functioning, role physical, bodily pain, and general health scales, whereas for the MCS summary score, higher weights are given to the vitality, social functioning, role emotional, and mental health scales.

Number of items. The SF-36 consists of 36 items, 35 of which are used in the calculation of 8 separate scale scores. The physical functioning scale (10 items) is the longest scale. The general health and mental health scales have 5 items each, and the vitality and role physical scales have 4 items each. The role emotional scale has 3 items, and the bodily pain and social functioning scales have 2 items each. The remaining item of the SF-36 is a health transition question that asks about a change in general health over the past 12 months.

The SF-12 consists of 12 items: 2 items on physical functioning, 2 items on role physical, 1 item on bodily pain, 1 item on general health, 1 item on vitality, 1 item on social functioning, 2 items on role emotional, and 2 items on mental health. Since more items permit better representation of each domain, the domains are best represented by the SF-36. The most useful measures derived from the SF-12 are the 2 aggregate summary measures: the PCS and MCS.

Response options/scale. The response scales for the SF-36 and SF-12 items vary across and within the scales, with the number of response options ranging from 3 (physical functioning) to 6 (vitality and mental health). The health transition item is scored on a 5-point scale where 1 indicates much better than a year ago, and 5 indicates much worse than a year ago.

Recall period for items. The SF-36 and SF-12 are available in 2 forms: a standard form, which uses a 4-week recall period, and an acute form, which uses a 1-week recall. The standard 4-week recall form is appropriate when the instrument will be administered only once to the respondent, or when at least 4 weeks will pass between re-administration of the instrument. The acute 1-week recall form is appropriate when more frequent administration is required and changes are likely to occur rapidly.

Examples of use. The SF-36 is ubiquitous in rheumatology and has been used to capture health-related outcomes in a variety of rheumatic conditions, including knee osteoarthritis (7), Sjögren's syndrome (8), fibromyalgia (9), rheumatoid arthritis (10,11), ankylosing spondylitis (12), and gout (13). The SF-36 has been used to assess efficacy of a broad range of interventions in rheumatology, including orthopedic surgery (14–16), drug treatment (8,17), acupuncture (18), physiotherapy (19), electromagnetic field therapy (20), Tai Chi (21), and self-management education (22).

The SF-12 has been used in population-based studies to assess the impact of musculoskeletal diseases on general health (23,24). In addition, it has been used as an outcome measure to evaluate the efficacy of a broad range of interventions for rheumatic conditions, including pharmacologic treatment (25,26); hydrotherapy treatment for osteoarthritis (27) and fibromyalgia (28); Tai Chi (29); surgical procedures (e.g., total hip arthroplasty) (30), fore foot arthroplasty (31); total knee arthroplasty (32); and medication adherence programs (33).

Practical Application

How to obtain. The original version of the SF-36 (Research and Development [RAND] 36-Item Health Survey 1.0 Questionnaire) can be obtained free of charge from the RAND Corporation (http://www.rand.org/health/surveys_tools/mos/mos_core_36item_survey.html). English and Arabic language versions are available. The revised SF-36 and the SF-12 can be obtained from QualityMetric (http://www.qualitymetric.com/). Annual license fee applies. License fees are available on application and depend on whether the survey is used in a commercial or nonprofit setting. Manuals can also be purchased.

Method of administration. The SF-36 and SF-12 can be self-administered or interviewer-administered. Multiple modes are available, such as static (paper), online, e-form, personal digital assistant, tablet, and interactive voice response (IVR) via telephone. Several studies reported a consistent bias for lower SF-36 and SF-12 scores (indicating worse health) when self-completed as compared with interviewer administration (34–38). For SF-36, data quality also tends to be better for interviewer administration with a lower proportion of missing data, lower ceiling effects, and better internal consistency estimates (35,39). Data collection costs, on the other hand, are lower (up to 77%) for self-administration (35,39). IVR and live telephone methods for administering the SF-12 have been compared in a study of back pain patients, with similar results obtained for PCS scores but not MCS scores (mean MCS 44.22 and 48.50 for IVR and live telephone methods, respectively; $P < 0.01$), and the greatest discrepancy occurring for the item about feeling "downhearted and blue" (40).

The SF-36 can also be administered by proxy, but concordance between self and proxy ratings varies across proxy types. Generally, professional proxies (e.g., occupational therapists, nurses) provide a more accurate description of an individual's health state compared with lay proxies, who tend to overestimate the level of impairment (41,42).

Scoring. The SF-36 and SF-12 contain a mixture of positively- (higher scores indicate better health) and negatively-worded response scales, so some items need to be recoded prior to scoring. The scale scores are calculated by summing responses across scale items and then transforming these raw scores to a 0–100 scale. Computerized scoring algorithms are available and can be used to produce norm-based T scores for each scale (with a mean of 50 and SD of 10) as well as the PCS and MCS summary scores (4). If using the IVR mode, data can be loaded directly into the QualityMetric database for scoring, interpretation, and reporting in real time.

In computing scale scores for the SF-36 and SF-12, missing values have traditionally been calculated only for those respondents who provided data on at least half the scale items (4). More recently, pattern matching and regression methods of missing data imputation for these questionnaires have been developed (43). These new algorithms can be implemented using QualityMetric's purpose-developed software.

Score interpretation. Scores on the SF-36 and SF-12 scales range from 0–100, with higher scores indicating better health. On the physical functioning scale, low scores are typical of someone who experiences many limitations in physical activities, including bathing or dressing, while high scores represent someone who is able to perform these types of activities without limitations. Low scores on the role physical scale represent someone who experiences many limitations in work or other daily activities, and high scores characterize someone who has no difficulties with these activities. Low scores on the social functioning typify a person who experiences a great deal of difficulties in normal social activities due to physical and emotional health problems, and high scores represent someone who is able to perform normal social activities without interference due to physical or emotional health. Low scores on the bodily pain scale are typical of a person who has very severe and extremely limiting pain, and high scores represent individuals who have no pain or pain-related limitations. On the mental health scale, low scores represent high levels of nervousness and depression, while high scores characterize someone who feels peaceful, happy, and calm. Low scores on the role emotional scale represent someone who experiences many problems with work or other daily activities as a result of emotional ill health, and high scores represent those who have no problems with work or other daily activities as a result of emotional health. On the vitality scale, low scores are typical of someone who feels tired and worn out all of the time, while high scores characterize those who feel full of pep and energy. Low scores on the general health scale represent a person who believes their health to be poor and likely to get worse, and high scores represent someone who sees their health as excellent (1).

Age- and sex-based norms for the SF-36 are available for several countries, including the US (4,44), the UK (34,45), Australia (6,46,47), Sweden (48), China (49), and New Zealand (50). Normative data for MCS and PCS summary scores are also available for Denmark, France, Germany, Italy, the Netherlands, Norway, and Spain (5). Notable cross-country differences in normative SF-36 scores have been reported (6), which may reflect cultural differences in health perceptions. Contextual factors, such as survey methodology, mode of administration, and item order have also been reported to affect normative scores on the SF-36 (34). Age- and sex-based norms for the SF-12 are also available for several countries, in particular the US (2,51,52). Unlike for the SF-36, SF-12 data from general population surveys in 9 European countries suggest there is little difference between standard US-derived scoring algorithms and country-specific algorithms, and standard scoring algorithms are recommended (53).

Respondent burden. The data on the respondent burden of the SF-36 are mixed. The self-reported version takes only 7–10 minutes to complete (54), although the presence of cognitive or physical impairment and depressed mood are associated with substantially longer completion time (55). The SF-12 takes only 2–3 minutes to complete (in a small pilot test, 81% completed the SF-12 in <2 minutes), less than one-third the time required to complete the SF-36 (2).

Generally, although the SF-36 and SF-12 use plain, easy-to-understand language, some of their items contain more than 1 concept (e.g., moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf), which could make it difficult for the participants to select the most appropriate answer. Evidence of item or response misinterpretation on the SF-36 has been reported in at least 2 studies (56,57).

Administrative burden. The SF-36 and SF-12 have relatively low administration burden. Interviewer administration of the SF-36 by telephone takes 16 to 17 minutes (58). No specific training for administration of the SF-36 or SF-12 is required, and completion instructions are self-explanatory. Computerized scoring algorithms for the revised versions of the SF-36 and SF-12 are available for purchase from QualityMetric and require basic knowledge of statistical software.

Translations/adaptations. The original versions of the SF-36 and SF-12 are available in English and Arabic. The revised versions are available in 121 languages. A list of translated versions is available at http://www.qualitymetric.com/WhatWeDo/LanguageTranslations/SurveysandTranslationsAvailable/tabid/215/Default.aspx, and further information can be obtained from the International Quality of Life Assessment web site (http://www.iqola.org). QualityMetric offers a translation service if required. Cultural adaptations of the original US version to other English-speaking countries are also available (59).

Psychometric Information

Method of development. The SF-36 was developed out of the RAND Corporation Health Insurance Experiment (60). The initial study measured 40 health concepts, 8 of which were selected for the inclusion into the new questionnaire. These 8 concepts were chosen to represent issues that were frequently used in health surveys and were affected by disease and treatment (3). Items for the questionnaire were generated following review of the content of various instruments that, at the time of SF-36 construction, were used to measure mental health and general health perceptions as well as limitations in physical, social, and emotional role functioning (1). No patient groups or representatives of the general population were involved in questionnaire construction. The SF-12 was developed using regression analysis methods to select and score 12 items from the SF-36 to reproduce the PCS and MCS scales in the general US population (2).

Acceptability. Data on acceptability of the SF-36 are mixed. The proportion of missing data varies from 0% for interviewer administration (61) to 26% for mailed versions (62) in nonhospitalized rheumatoid arthritis patients to 47% in hospitalized patients (including musculoskeletal patients) (63). Higher proportion of missing data is significantly associated with increasing age and disability (55,63).

On the SF-12, missing data in rheumatology settings occur less frequently, with just 15% of Danish respondents with arthritis missing ≥1 items of the PCS and 16% missing ≥1 items of the MCS (64). There is also a low individual item missing rate (<2.30%) and high percentage score computability (>90%) (25).

In arthritis studies, ceiling effects (>10% of participants obtaining the lowest possible score) are commonly reported for SF-36 role physical (21–76%), role emotional (49–60%), social functioning (23–64%), and bodily pain (20–40%) scales (14,61,65). Ceiling effects have also been observed on the mental health (20–28%) (14,65) and vitality (18%) scales (65). Floor effects (>10% of participants obtaining the highest possible score) are frequently found on role physical (29–80%) and role emotional scales (27–48%) (14,61,62,66,67,68), while at least 1 study has found there to be no notable floor effects (65). There do not appear to be ceiling and floor effects for the SF-12 among patients with rheumatic conditions (25,64).

Reliability. In musculoskeletal settings, results for reliability of the SF-36 are mixed. In several studies, all of the SF-36 scales were reported to have good internal consistency, with Cronbach's $\alpha \geq 0.70$ (61,62,65,67,69). In addition, internal consistency estimates were in excess of 0.90 for physical functioning and bodily pain scales in at least 2 studies (61,62) and for general health in at least 1 study (70), indicating suitability of these scales for use at the level of individual. Results for test–retest reliability of the SF-36 are less encouraging with the intraclass correlation coefficient (ICC) below the recommended standard of 0.70 on mental health (ICC 0.55) (68), role emotional (ICC 0.66) (68,71), as well as role physical (ICC 0.44), and vitality (ICC 0.03) scales (71).

Evidence also indicates a high proportion of measurement error on the SF-36 questionnaire in rheumatology, with large SDs for one-time administration (up to or exceeding the mean score) (72) and large variations (change of up to 200% from the initial score) in the SF-36 scores over a one-week test–retest period (70–72). In orthopedic surgery, minimal detectable change at an individual level ranged from 22% (general health) to 97% (role physical) of the total score range (14,70).

Internal consistency of the SF-12 component summary scores is generally high (Cronbach's $\alpha \geq 0.82$ and 0.75, for SF-12 PCS scale and MCS scale, respectively) (73–76). Test–retest reliability of the SF-12 administered 2 weeks apart is adequate in the US and the UK general populations: $r = 0.89$ for PCS and $r = 0.76$ for MCS (2), and others (74,77,78). For both the PCS and MCS scales, changes in scores between test and retest averaged less than 1 point, and at the second administration, 85% scored within the 95% confidence interval (95% CI) of the score at the first administration (2).

Validity. *SF-36.* Given the uncertainty about the reliability of the SF-36, findings on the validity of this measure need to be interpreted with caution. The SF-36 appears to have good face validity, with all items referring to health-related issues. Although content validity of an outcome measure is largely determined by the concept being measured, and may vary from one setting to another, the SF-36 captures a broad range of health states. However, the presence of severe floor and/or ceiling effects on the number of the SF-36 scales in rheumatic conditions indicates that this questionnaire does not adequately capture the full range of health experiences in this setting. Empirical studies of the construct validity of the SF-36 have shown mixed results for its validity.

The dimensional structure of the SF-36 (8 first-order and 2 higher-order factors) has been questioned in several studies. For example, higher-order factor analysis of the scale scores have confirmed separation of the scale scores into mental and physical health summary scores in some rheumatic studies (62,69) but not others (79). First-order factor analysis has also failed to confirm the 8-dimensional structure of the SF-36 in either exploratory (61,69) or confirmatory factor analysis (79,80).

Results for the convergent validity of the SF-36 are generally favorable. In several studies (62,65,70,72), the SF-36 scales had higher correlations with measures of similar constructs (such as the Western Ontario and McMaster Universities Osteoarthritis Index, the Nottingham Health Profile, the Health Assessment Questionnaire [HAQ], and rheumatoid arthritis disease activity measures) and lower correlations with dissimilar domains. In at least 2 other studies, the SF-36 had expected strong correlations ($r > 0.60$) with measures of similar concepts, including the Arthritis Impact Measurement Scales (AIMS), the HAQ, and EuroQol 5-domain instrument (EQ-5D) (67,68). However, at least one of these studies reported that the SF-36 scales did not correlate as well as expected with disease-specific measures of rheumatoid arthritis (61). However, the discriminant validity of the SF-36 has received less support. In a study of 200 patients with rheumatoid arthritis, physical functioning, role physical, general health, and bodily pain scales were expected to have correlations <0.3 with questionnaires hypothesized to measure dissimilar concepts (including fatigue and rheumatoid arthritis disease activity) (68). The discriminant correlations were much higher than expected, ranging between 0.51 (correlation between physical functioning and visual analog scale of fatigue) and 0.78 (correlations between bodily pain and rheumatoid arthritis activity scale). Higher than expected correlations between SF-36 scales and conceptually dissimilar measures were reported in at least one other rheumatoid arthritis study (67).

Overall, the evidence supports the known-groups validity of the SF-36 in rheumatology. With the exception of the mental health scale, the SF-36 had been able to differentiate between levels of osteoarthritis severity (72). The difference between those who had moderate and severe osteoarthritis, assessed using standardized effect sizes (ES), were in the small to moderate range, varying from 0.35 for general health to 0.75 for physical functioning. In the same study, all scales but role emotional and pain were able to differentiate among rheumatology patients with and without comorbid conditions, also with small to moderate ES, ranging from 0.49 (physical functioning and mental health) to 0.78 (general health). In another study, the SF-36 physical functioning and bodily pain scales discriminated well between patients receiving the disability pension versus those who did not, with medium ES values (0.69 and 0.50, respectively) (68). The SF-36 has also been shown to be able to differentiate between people with and without lower extremity osteoarthritis (81).

SF-12. Given that the primary purpose of the SF-12 was to reproduce the PCS and MCS scores of the SF-36, how well it does so is the important criterion, and there is strong evidence for the criterion-related validity of the SF-12. The SF-12 PCS and MCS scores correlate 0.95 and 0.96 with the SF-36 PCS and MCS scores, respectively (2,53). These findings have been replicated in the general populations of 9 European countries (Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, and the UK) (53), with very high correlations between SF-12 PCS and SF-36 PCS scores ($r = 0.94-0.96$) and SF-12 MCS and SF-36 MCS scores ($r = 0.94-0.97$). Clinical trials data from patients with osteoarthritis and rheumatoid arthritis also indicate good criterion validity of the SF-12 in rheumatology, with strong correlations between the SF-12 and SF-36 PCS scores and the SF-12 and SF-36 MCS scores ($r = 0.92–0.96$) (25).

The 2-factor conceptual structure of the SF-12 (PCS and MCS) has been confirmed in several population-based (82,83) and clinical studies (25,84). However, a recent study has challenged the 2-factor structure of the SF-12 (73). The standard orthogonally-weighted SF-12 scoring algorithm has been cautioned against, with oblique scoring algorithms appearing preferable (73,85).

Convergent and discriminant validity of the SF-12 in a general population is supported by relationships found with the EQ-5D (82). Comparable summary scores and dimensions correlate better, e.g., PCS with mobility ($r = -0.69$), usual activities ($r = -0.71$), and pain discomfort ($r = -0.61$) and MCS with anxiety/depression ($r = -0.47$), indicating good convergent validity. Less comparable summary scores and dimensions correlate weakly, e.g., PCS and anxiety/depression ($r = -0.28$) and MCS and mobility ($r = -0.34$), supporting discriminant validity of the SF-12.

Results for the convergent and discriminant validity of the SF-12 in rheumatic diseases are somewhat variable. In Danish patients with rheumatoid arthritis (64), the SF-12 PCS and MCS have been found to have unexpectedly weak correlations with measures of similar constructs, such as the HAQ ($r = -0.15$ for PCS and $r = -0.25$ for MCS) and lower correlations with dissimilar domains. In spinal clinic patients, back pain and disability have been found to be significantly moderately correlated with SF-12 PCS ($r = -0.41$ and -0.63, respectively, $P < 0.0001$) and MCS ($r = -0.33$ and -0.55, respectively, $P < 0.0001$) (76), as hypothesized. In addition, as expected in these patients, stress has been found to be weakly correlated with SF-12 PCS ($r = -0.07$, $P = 0.001$), but moderately correlated with SF-12 MCS ($r = -0.33$, $P < 0.0001$) (76).

The ability of the SF-12 to differentiate between groups based on severity of health impairment is generally good and is similar to that of the SF-36. PCS-12 and MCS-12 reach the same statistical conclusions about group differences as PCS-36 and MCS-36; they do so with relative validity coefficients that are typically 10% below those observed for the SF-36 (2). More specifically, among the Greek general population, the SF-12 PCS score has been found to be significantly worse among those reporting hip and knee problems compared with those not reporting such problems ($P < 0.01$) (82). However, among Danish patients with rheumatoid arthritis, the SF-12 did not seem to be sensitive to variation between patient groups with different disease severity (64).

Ability to detect change. Systematic examinations of minimum clinically important differences (MCID) for the SF-36 in rheumatic conditions are rare. More generally, a minimal detectable change of 5 points on a 100-point scale was previously reported for the SF-36 and is based on 95% CIs from a normative sample (4). Other measures of responsiveness (standardized response mean [SRM] and ES) support the ability of the SF-36 to detect change in this setting. The SF-36 has been demonstrated to be able to detect large improvements in health status at 3 and 6 months following joint replacement surgery (81). The most responsive scales were physical functioning, role physical, bodily pain, and social functioning, with SRMs of 1.04 or higher at 3 months. The least responsive scales were general health (SRM 0.20) and role emotional (SRM 0.37). Similar results have been obtained in another joint replacement study, with large improvements (ES >0.80) recorded for physical functioning, role physical, and bodily pain, moderate improvements (ES 0.50–0.80) for role emotional, vitality, and social functioning, and small improvements (ES <0.50) for mental health and general health at 6-months followup (14).

Veehof compared the responsiveness of the SF-36 with the responsiveness of disease specific scales (including the AIMS2 and the HAQ) in 168 patients with rheumatoid arthritis (86). The study showed that the responsiveness of the SF-36 is comparable to that of disease-specific measures. The bodily pain, vitality, physical functioning, and role physical scales have also been shown to have good ability (assessed by SRM) to identify rheumatoid arthritis patients who were classified as improved based on self-rating of disease activity (61,68). Responsiveness of the SF-36 to deterioration is more limited, with no scale able to capture self-reported deterioration (68).

The MCID of the SF-12 in rheumatology are also not known and there is limited information on its responsiveness. There is some evidence among those with back pain attending a spinal clinic (76), with a large ES for SF-12 PCS (0.82) and a small to moderate ES for SF-12 MCS (0.37) observed in patients whose self-reported back pain became much better after 3–6 months of followup. Similarly, moderate ES for SF-12 PCS and MCS (−0.46 and −0.21, respectively) were observed in patients whose self-reported back pain became much worse. Among workers with neck or upper extremity musculoskeletal disorders, SF-12 PCS scores have been shown to be responsive to clinically confirmed incident cases with a decrease in general physical function observed (ES −0.9, SRM −0.6). SF-12 MCS has been shown to be not responsive to such change, and neither PCS nor MCS scores were responsive to self-reported symptomatic incident cases, self-reported symptomatic recovered cases, or clinically-confirmed recovered cases (ES or SRM <0.2 or changes not in the expected direction) (87). In addition, SF-12 was reported to be responsive to a wide range of treatments and programs for musculoskeletal diseases (26–29,33).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SF-36 and SF-12 can be used when the assessment of a broad range of health aspects is needed. The SF-12 is brief, appears to adequately reproduce the 2 summary scores of the SF-36, and generally has comparable psychometric properties to the SF-36. Since the concepts represented in these questionnaires are not disease specific, the SF-36 and SF-12 are especially suited when comparisons between disease groups or with the general population are required. Availability of population norms also provides context for score interpretation. The SF-36 and SF-12 also appear to differentiate between levels of disease severity in rheumatic conditions and between people with and without rheumatic conditions, as well as respond to treatment-related changes in health status of people with rheumatic conditions.

Caveats and cautions. Given equivocal evidence of psychometric robustness of the SF-36 in rheumatic conditions, its use in this setting needs to be approached with caution. The role physical, role emotional, and social functioning scales are frequently reported to have low reliability, which puts their validity into question. Large test–

retest variations in the SF-36 scores at the individual level make the SF-36 unsuitable for individual assessments. Floor and ceiling effects in rheumatic conditions also indicate that the SF-36 does not adequately target the full range of health experiences of this population.

In contrast to the extensive SF-36 literature, the SF-12 has been less well-studied. Findings related to the SF-36 may not be transferable to the SF-12. There is a small loss (10%) in the ability of the SF-12 to distinguish between different disease groups compared with the SF-36. Use of the SF-12 for assessing and/or monitoring individuals is discouraged. There is limited evidence of its responsiveness to treatment-related changes in the health status of people with musculoskeletal conditions.

Clinical/research usability. In clinical settings, large intra-individual variations in the SF-36 scale scores and its low ability to detect deterioration make it unsuitable for use with individual patients, although the scale appears to have satisfactory ability to detect treatment-related improvements in health at a group level. In research settings, the SF-36 can be used to compare different disease groups or disease groups with the general population. Low measurement precision reported for the SF-36 scales in cross-sectional and test–retest studies can also dramatically increase sample size required to detect the desired ES for either between-group differences or within-group change over time. Ease of administration, availability of an online version, and availability of a computerized scoring algorithm support the usability of the SF-36 and SF-12 in research settings. However, financial costs can limit the use in low-budget studies, although the original version of the SF-36 is available at no cost. The SF-12 is a suitable measure where information on the SF-36 PCS and MCS scores is required. Psychometric evaluation does not support interpretation of scores to make decisions for individuals and, therefore, limits its clinical use.

NOTTINGHAM HEALTH PROFILE (NHP)

Description

Purpose. The NHP was developed in the 1970s (88) for measuring the impact of illness on patients and the assessment of changes in health status over time (89). As a generic health status questionnaire, it provides a brief indication of a patient's perceived emotional, social, and physical health and is intended for use in the general population (90).

Content. There are 2 parts of the NHP. The domains covered in part 1 are related to the health status of the individual (89) and include energy levels, pain, emotional reactions, sleep, social isolation, and physical abilities. Part 2 addresses the impact of ill health on daily life (89) and covers paid employment, home duties, social life, home life (relationships), sex life, interests and hobbies, and vacations. The 2 parts of the NHP can be used together or separately, with part 1 frequently used on its own.

Number of items. Part 1 consists of 38 items. The energy levels domain has 3 items, pain has 8 items, emotional reactions consists of 9 items, sleep and social isolation domains have 5 items each, and physical abilities domain has 8 items. Part 2 consists of 7 items that cover the 7 life areas listed in the above section (91).

Response options/scale. Responses are measured on a dichotomous scale, with respondents asked to check a yes box or a no box, according to whether a statement applies to them. If unsure, the instructions are to select an answer that is more applicable at the time of answering the questionnaire.

Recall period for items. Respondents are asked to identify whether each statement applies to them "at the moment."

Examples of use. In rheumatology, the NHP has been used in several randomized controlled trials, including evaluation of outcomes of exercise programs in rheumatoid arthritis (92), manual lymph drainage therapy and connective tissue massage in primary fibromyalgia (93), balneotherapy and tap water (94), and balneotherapy and mud-pack therapy (95) in patients with knee osteoarthritis. The NHP has also been used in several observational studies for the purpose of evaluating health status of people with osteoarthritis after knee arthroplasty (96) and after hip revision surgery (97), assessment of the efficacy of disease-modifying antirheumatic drugs in rheumatoid arthritis (98), and to evaluate outcomes of multidimensional rehabilitation program in chronic myofascial pain and/or fibromyalgia (99).

Practical Application

How to obtain. A copy of the NHP can be viewed at www.cebp.nl/media/m83.pdf. The official web site (www.galen-research.com) was under development at the time of writing; however, a copy of the NHP can also be obtained by contacting Galen Research (gr@galen-research.com). The noncommercial license fee for one language version of the NHP is approximately $192 (£120) per study. The scoring manual is included in this cost. The minimum cost for commercial studies is approximately $8,000 (£5,000) for 1 language version and increases to approximately $24,000 (£15,000) for 2 languages with an additional $8,000 (£5,000) for each subsequent language.

Method of administration. The NHP is designed to be self-administered.

Scoring. A scoring algorithm is available with the purchase of the questionnaire. Scoring instructions can also be downloaded from https://www.cebp.nl/media/m83.pdf. Scores for each of the 6 domains in part 1 are computed by summing weighted values given to each positive response. The weights for the NHP were derived using Thurstone's method of paired comparisons from a sample of 215 members of the general public. The sum total of the weighted scores is 100, with weights intended to reflect the perceived severity of a health state represented by the item from the point of view of the general public, rather than a specific patient population (89). Only domain scores are calculated, with no overall score.

There appear to be no specific instructions for handling missing values. Developers of the NHP recommend scoring responses to missing items as "no" since the respondents did not answer "yes" to these questions. However, Kersten et al (100) caution against using this approach routinely, since it could substantially underestimate the level of disability, particularly for severely disabled people, such as those using wheelchairs who are unable to walk at all.

Score interpretation. The scores on NHP domains range from 0 (best health state) to 100 (worst health state) (91). No normative data for the NHP are available.

Respondent burden. The NHP appears to have low respondent burden, taking 5–10 minutes to complete (91). Developers of the NHP have described this questionnaire as being a simple instrument that is acceptable and understood by a majority of people (91). In general, statements in the NHP are simple and easy to understand; for example, "I feel lonely" or "I have pain at night." However, some statements describing negative health states (e.g., "I feel that life is not worth living") may distress some respondents.

Administrative burden. Scoring of part 1 produces 6 domain scores plus a further 7 scores are produced if part 2 is used, therefore scoring may be cumbersome if done by hand (91). However, scoring and administration instructions are self-explanatory and require no specific training.

Translations/adaptations. The NHP is available in numerous languages including English, Greek (101), French (102), Swedish (103), Dutch (104), and Spanish (105). For an extended list of available translations, see http://www.proqolid.org/instruments/nottingham_health_profile_nhp.

Psychometric Information

Method of development. Information on the development of the NHP part 1 is generally scant and lacking in detail. In the development of part 1, statements describing the typical effects of ill health (social, psychological, behavioral, and physical) were collected from more than 700 people (91). This initial stage produced 2,200 statements, with 138 statements left after the removal of redundant and ambiguous items. The properties of these 138 statements were evaluated in a number of studies using diverse patient populations, after which the number of statements was reduced to 82 (91). No further information on characteristics of study participants, or types of tests or criteria used in item refinement and selection, is provided in the original publication describing the development of the NHP.

Part 2 was subsequently developed for the purpose of assessing how perceived health problems may affect daily living (106). The original statements collected during the development of the NHP were reviewed to identify areas of "task performance" most often affected by health problems. The areas of job, housework, social life, family life, sex, spare time activities, holidays, and travel were identified. Interviews were conducted with patients attending a hospital outpatient clinic. Difficulties in wording and presentation were identified, and further interviews were conducted with outpatients and a range of university employees. In total, 114 interviews were conducted. The wording of the items was revised by the developers with the intent of making them more understandable and acceptable for the average person with no university background and possibly limited education (106).

Acceptability. Missing data may be an issue when the NHP is administered to people who are severely disabled. In a study of 92 people with a range of disabilities (including 7 with rheumatoid arthritis), 46 people were unable to complete the NHP due to questions referring to activities that they were unable to perform (100). Missing data were present on 14 of 38 (37%) questions. Questions relating to pain, standing, walking, and other physical activities such as climbing stairs were particularly problematic. The pain domain was not completed by 48% of participants, and the physical functioning domain was not completed by 49%.

The NHP appears to be better able to capture states of ill health rather than states of good health. More than 50% of respondents in a study comparing the NHP sores for consulters of a general practice and nonconsulters scored 0 (best health) on each of the NHP domains (90). In a more recent study of 111 people using wheelchairs who live independently (including 30 who had rheumatic conditions), the emotional reactions, social isolation, and sleep scales of the NHP all had median scores of 0 (107).

Reliability. A limited number of studies have examined the internal consistency of the NHP in rheumatic conditions and have reported mixed results. The internal consistency of the NHP pain subscale was found to be acceptable in a sample of 160 people with rheumatoid arthritis (Cronbach's α = 0.83) (108). In another study conducted with a sample of 111 wheelchair-using people with a range of chronic conditions (including rheumatic diseases), the internal consistency of the pain and emotional reactions subscales were similar (Cronbach's α = 0.82), although the internal consistency of the mobility subscale was very poor (Cronbach's α = 0.34), and no internal consistency estimates were reported for the remaining subscales (107). In a sample of 1,063 individuals drawn from the general population, the internal consistency of the social isolation subscale was slightly below the acceptable lower limit of 0.70 (Cronbach's α = 0.65) while the internal consistency of the remaining subscales ranged from 0.71 (energy) to 0.88 (pain) (109).

Information on test–retest reliability of the NHP in rheumatology settings is very limited. In a sample of 73 patients with osteoarthritis who had no other comorbidities, 4-week test–retest reliability of the NHP (assessed using Pearson's correlation coefficient) ranged from 0.77 (energy) to 0.85 (sleep and physical mobility) on part 1 and 0.44 (hobbies/interests) to 0.86 (paid employment) on part 2 (110). However, it is well recognized that Pearson's correlation coefficient is a poor measure of temporal stability, since it is unable to capture systematic changes in scores over time. Hence, the "real" test–retest reliability of the NHP might be even lower. In a sample of 49 individuals with musculoskeletal disorders, test–retest reliability (assessed using intraclass correlation coefficient [ICC]) was within the acceptable range for pain (ICC 0.87) and physical ability (ICC 0.76) scales, with no information provided for the remaining NHP scales (111). Test–retest reliability of the NHP subscales in a French study, conducted with 111 individuals with rheumatoid arthritis, of the NHP subscales ranged from 0.57–0.73 (112).

Validity. The NHP also appears to have good face validity, with all items referring to an aspect of health. Although content validity of a measure is largely dependent on the concept being measured, the NHP covers a broad range of health-related functions (physical abilities, pain, sleep) that could be expected to be affected in rheumatic as well as in many other chronic health conditions, and therefore appears to have good content validity as a measure of general health status. In the development of the NHP, patient consultation was combined with expert consultation, therefore enhancing the relevance of the questionnaire to patients and clinicians. However, there has been limited investigation of the factor structure of the NHP in rheumatology-specific populations and more generally, so very little is currently known about the factorial validity of this questionnaire.

Convergent and discriminant validity of the NHP in rheumatology settings appears to be supported. A 3-year followup study of people with rheumatoid arthritis (n = 160 at baseline, n = 124 at 3-year followup) found that correlations between the pain subscale of the NHP and the General Health Questionnaire-28 ranged from 0.45 and 0.64, and from 0.25 and 0.41 between the pain subscale of the NHP and the Ritchie Articular Index (108). Less pain was also significantly correlated with greater psychological well-being. This profile of associations was in line with author hypotheses. Similarly, in a sample of 72 individuals with rheumatoid arthritis, the NHP showed an expected pattern of correlations with the Arthritis Impact Measurement Scales (AIMS), Beck Depression Inventory, and Health Assessment Questionnaire (HAQ) (113). In another study, all 6 NHP domain scores were significantly ($P < 0.0005$) related to disease activity measured by the Modified Disease Activity Score in rheumatoid arthritis, ranging from 0.25 for social isolation to 0.55 for pain (114).

Known-groups validity of the NHP was assessed in several studies and also received robust support. The NHP was able to differentiate between people with rheumatoid arthritis and a sample of well, community-dwelling people ages 40–59 years, with significantly lower scores for the rheumatoid arthritis group on the domains of energy, pain, physical mobility, and sleep (113). Scores for emotion and social subscales of the NHP were more similar between these groups, although mean scores were poorer in the rheumatoid arthritis sample for each domain. In another study, 200 outpatients with rheumatoid arthritis had higher NHP scores than both a random population sample and a second sample of patients with a variety of common diseases (114). However, neither of the above 2 studies provided standardized measures of differences, therefore information on magnitude of the differences in NHP scores between people with and without rheumatic conditions awaits further research.

The NHP also appears to be able to differentiate people with rheumatic conditions from those with other types of chronic illness. In a study of 82 individuals with rheumatoid arthritis or migraine (89), the authors hypothesized that rheumatoid arthritis would have greater impact on individual health than migraine, which would be reflected in higher NHP domain scores for individuals with rheumatoid arthritis. This hypothesis was partially supported. People with rheumatoid arthritis did have significantly worse health than the migraine group, but only on 3 out of the 6 NHP domains, including energy (migraine 43.6, rheumatoid arthritis 74.3), pain (migraine 14.9, rheumatoid arthritis 67.3) and physical mobility (migraine 2.0, rheumatoid arthritis 64.6), with no significant differences between the groups on domains of sleep, emotional reactions, and social support (89).

Ability to detect change. Information about the ability of NHP to detect change in rheumatic conditions is somewhat inconsistent, although it generally indicates that the NHP may not be as sensitive to change as other instruments that measure similar concepts. In one study, the ability of the NHP to detect self-reported improvements in health status in rheumatoid arthritis was compared with that of the AIMS, the HAQ and the Functional Limitations Profile (FLP) (115). Not one instrument outperformed the others across all domains. Compared with other questionnaires, the NHP had the lowest ability to detect self-reported change in mobility (effect size [ES] 0.27), pain (ES 0.38), and emotion (ES 0.59) domains, with only small to moderate ES recorded. At the same time, ES for other questionnaires measuring similar concepts were in moderate to high range, ranging from 0.69 to 0.83. In the social domain, NHP (ES 0.24) was worse at detecting change than FLP (ES 0.60) but better than the AIMS (ES 0.06). In another study involving 276 people with unilateral osteoarthritis of the hip waiting for joint replacement surgery, NHP was able to detect change in health status, with all NHP domain scores showing significant improvements 1 year following the surgery (116), although no information on the magnitude of change had been reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The NHP encompasses several domains that are of relevance to rheumatology, including energy, pain, physical mobility, emotions, sleep, and social and holds several areas in common with other disease specific instruments in this field. Being a generic measure, the main advantage of the NHP compared with disease-specific measures is that it can be used to compare the impact of rheumatic conditions with that of other illnesses or with the general population.

Caveats and cautions. Although studies assessing construct validity of the NHP produced favorable results, low test–retest reliability of some domains (namely emotion and social) raise doubts about psychometric robustness of this measure in rheumatology. It would also appear that the NHP is not appropriate for use in people with minor disability due to severe ceiling effects. The presence of ceiling effects could also pose problems in pre- and post-intervention studies, since improvement in condition for those who score zero at baseline cannot be demonstrated. Furthermore, the NHP also appears to be less sensitive to change than other health status measures used in rheumatology. The use of the NHP with severely disabled people might also present problems due to large amounts of missing data (100), and there appears to be no adequate methods for handling missing data on this questionnaire.

Clinical/research usability. The NHP is easy to use and score. Part 1 contains several items within each domain that combine to form a moderately detailed picture of the patient's current health. The areas of life affected by health listed in part 2 could serve to flag areas for further assessment in a clinical context. However, the high cost of obtaining the questionnaire could limit its usefulness in clinical settings. Sensitivity to change of the NHP is lower than that of other instruments measuring similar aspects of health, therefore its use in clinical trials where ES largely dictates the size and cost of a trial is less assured. As a measure of general health and health-related quality of life, the NHP may be of greater interest in epidemiologic rather than clinical research, although pronounced ceiling effects in well populations may severely limit its usefulness in population-based studies.

SICKNESS IMPACT PROFILE (SIP)

Description

Purpose. The SIP is a generic measure of health-related functional status (117), designed to be broadly applicable across types and severities of illness and across demographically and culturally diverse groups (118). The purpose of the scale is to provide a descriptive profile of changes in a person's behavior due to sickness (119). The SIP is intended for use in health surveys, program planning, policy formation, and monitoring patients' progress (120), and was initially published in 1977. Due to the length and respondent burden of the original 136-item version of the SIP, a shorter version, the SIP68, with 68 items, was developed in 1994 (121). A number of disease-specific short-form adaptations of the SIP were also developed, including back pain (122) and rheumatoid arthritis (123) versions.

Content/number of items. This SIP136 has 12 domains addressing the impact of health on a range of day-to-day behaviors, including sleep and rest (7 items related to sleep quality and daytime tiredness), emotional behavior (9 items addressing emotional well-being), body care and movement (23 items related to self-care, balance, and body movement), household management (10 items related to activities of daily life), mobility (10 items related to the ability to move within and outside the home), social interaction (20 items addressing relationships with others), ambulation (12 items related to walking), alertness behavior (10 items describing alertness and ability to concentrate), communication (9 items related to spoken and written communication), work (9 items related to work productivity and relationships with coworkers), recreation and pastimes (8 items addressing frequency and type of recreational activities performed), and eating (9 items addressing quantity and type of food intake) (120).

The SIP68 has 68 items across 6 domains: somatic autonomy (17 items related to basic somatic functions, such as ability to move independently and self-care), mobility control (12 items related to walking and hand use), psychic autonomy and communication (11 items describing concentration and spoken and written communication), social behavior (12 items related to social activities and recreation), emotional stability (6 items related to emotional self-control), and mobility range (10 items related to tasks of daily life) (118). The dimension names of the SIP68 differ from those of the SIP136, since during the construction of the SIP68, the questionnaire items formed a configuration of factor loadings that was different from that originally reported for the SIP136, with somewhat different dimensions emerging.

Response options/scale. When answering the SIP, respondents are asked to check all the statements that apply to them. Statements that do not apply are left blank.

Recall period for items. The recall period for all items is "today."

Examples of use. In rheumatology, the SIP has been previously used to assess changes in health-related function status following total hip replacement surgery (124), to determine the effects of an exercise program in osteoarthritis of the hip (125), and to measure the impact of multidisciplinary team care versus regular outpatient clinic care on overall health in people with rheumatoid arthritis (126).

Practical Application

How to obtain. A copy of the SIP136 is available under a limited use agreement from MAPI Research Trust at http://www.mapi-trust.org/questionnaires/53. The SIP costs ~$677 (€500) per study for funded academic research and ~$1,354 (€1,000) per study for commercial studies. There are no distribution fees for nonfunded academic research and individual clinical practice. Distribution fees are ~$400 (€300) per study, plus $68 (€50) per language version in funded academic research, and ~$677 (€500) per study plus ~$203 (€150) per language version in commercial studies (127). The SIP68 is available at no cost in de Bruin et al (121).

Method of administration. The scale can be self-administered or interviewer-administered.

Scoring. SIP scores can be calculated manually or using a scoring algorithm, which is available with the purchase of the SIP (128). The 12 categories of the SIP136 can be scored separately to provide a health profile. Alternatively, the SIP items can be combined to obtain 2 summary dimension scores, including physical dimension (ambulation, mobility, and body care/movement) and psychosocial dimension (emotional behavior, alertness behavior, communication, and social interaction) scores. An overall score based on all 136 items can also be obtained (120).

The category scores are calculated by adding the weights assigned to each item checked within the category. The sum total is then divided by the value of the highest weight for the category and multiplied by 100 to obtain the category score. The 2 dimension scores and the overall score are calculated in a similar manner. The item severity weights for the SIP have been derived using equal-appearing interval scaling method from a sample of more than 100 judges, including patients and health professionals in Seattle, Washington (120).

The SIP68 can be used to calculate an overall total score, 2 dimension scores (physical and psychosocial), or 6 subscale scores. The physical dimension score includes somatic autonomy, mobility control, and mobility range scales, and the psychosocial dimension consists of psychological attention and communication, social behavior, and emotional stability scales. The SIP68 is scored by adding the number of items that were checked for each category, dimension, or overall to obtain category, dimension, and total score, respectively. In scoring of SIP instruments, all unchecked items are given a score of 0.

Score interpretation. The score range for the SIP136 category, dimension, and total scores is 0 (best health) to 100 (worst health) (128). The score range for the SIP68 is 0 (best health) to 68 (worst health), with the score range for the category and dimension scores varying according to the number of items that make up a given category/dimension (118).

Respondent burden. The SIP136 may have moderate respondent burden, with an average completion time of 20–30 minutes (129). One study reported that 84% of respondents self-completed the SIP in <40 minutes (130). In a study of 168 male veterans residing in nursing homes in the US, the interviewer-administered SIP136 completion time ranged from 20–65 minutes, with a mean of 35 minutes. Longer completion times were associated with impaired verbal functioning. Interviewer assessment indicated that, in general, the instructions were well understood and items were not considered to be unduly sensitive. The SIP68 has been reported to take 15–20 minutes to complete (132).

Administrative burden. The SIP questionnaires have minimal administrative burden. The manual scoring procedure has been reported to take 5–10 minutes to complete for the SIP136. No special training is needed to either administer the questionnaire or interpret the results (129). Administration and scoring instructions are self-explanatory and are easy to follow for the SIP68.

Translations/adaptations. The original language of the SIP136 is US English. Existing translations (which may not have undergone a full linguistic validation process) are available in Arabic, Chinese for Hong-Kong, Danish, Dutch, Dutch for Belgium, English for Mexico, English for the UK, Finnish, French, French for Belgium, German, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Spanish for Mexico, Spanish for the US, Swedish, Tamil, and Thai (133). A UK adaptation of the SIP, the Functional Limitations Profile, is also available (134). Since the SIP68 is a shortened version of SIP136, this questionnaire can also be made readily available in multiple language versions.

Psychometric Information

Method of development. Statements describing sickness-related changes in behavior were elicited from general practice patients, health care professionals, significant others, and apparently healthy individuals (135). A total of 1,100 responses to the survey were collected. These statements, together with a review of function assessment instruments designed for the evaluation of circumscribed patient groups, resulted in 1,250 specific statements of behavioral change. These statements were subjected to standard grouping techniques according to a set of criteria, which yielded 312 unique statements, each describing a behavior or activity and specifying a dysfunction. A standard sorting procedure yielded 14 groups of statements, each of which appears to describe dysfunction in an area of living or a type of activity (119). The 14-item groups were further refined to produce the current scale with 12 domains.

The SIP68 was developed using principal components analysis of the data obtained from studies in 10 different diagnostic groups with a total of 2,527 respondents to the Dutch translation of the original SIP (121). Of the 2,527 respondents, data from 835 individuals were used in the construction of the SIP68, with a maximum of 100 from any of the 10 diagnostic groups (n = 100 for rheumatoid arthritis, n = 100 for ankylosing spondylitis, n = 41 for spinal cord injury, n = 53 for stroke, n = 100 for cancer, n = 100 for neuromuscular disease, n = 100 for back/neck pain, n = 100 for head injury, n = 99 for hemodialysis, and n = 42 for Crohn's disease). Items which applied to <10% or >90% of any diagnostic subpopulation were removed, as were items that did not contribute substantially (using an a priori definition of substantial loading as >0.40) to the scales or the total score.

Acceptability. The evidence for acceptability of the SIP instruments is generally not favorable. In a pilot study of the interviewer-administered SIP136 to 246 general practice enrollees (inpatients, home care patients, walk-in patients, outpatients, and nonpatients) in the US, all subjects completed the interview, with 9% of participants not finding at least 1 item on the questionnaire that applied to them (119). Similarly, in a study of 85 people with rheumatic conditions, who consented to participate in an evaluation study of the SIP68 in the Netherlands, 9% were unable to complete the instrument due to physical limitations or difficulty in understanding the instructions (118).

The proportion of missing data on the SIP is difficult to estimate, due to respondents instructed to leave items that do not apply to them unchecked. In a study of 301 people age ≥65 years, the question asking about sexual activity was left unchecked most frequently (12% of respondents) (136). In another study of 329 poststroke patients who participated in the interviewer-administered SIP136, responses from only 10 people (3%) could not be used for data analysis due to high proportion of missing data (137). While this study was not conducted in a rheumatology specific population, results may be indicative of a broader acceptability among frailer populations (121).

The SIP appears to have good range of functioning at the levels of very good health, with no floor effects generally reported for the total scale or dimension and category scores (136,138). However, substantial ceiling effects were found for category scores on the SIP136 in a study of 301 people, age ≥65 years, ranging from 31% for social interaction scale to 87% for the work scale. It should also be noted that persons who do not work at all (e.g., retired individuals) are given the maximum score for this category, therefore potentially inflating the ceiling effect for this scale. The physical and psychosocial dimension scores also had ceiling effects, with 27% and 22% of respondents recording best possible health state, respectively (136). These results indicate that the SIP may not be suitable to use with people who have low to moderate levels of ill health.

Ceiling effects were also reported for the SIP68. In a study of 329 people with disabilities (138), ceiling effects were found on emotional stability (54%), mobility range (24%), psychic autonomy and communication (24%), and somatic autonomy (17%) categories, as well as psychological dimension (19%). De Bruin et al (121) also found mild ceiling effect for the SIP68 total score in a sample of 83 outpatients with rheumatoid arthritis (12%). These individuals judged their health as good to very good, which was matched by a rheumatologist's rating of their functional status.

Reliability. In a sample of 299 patients with musculoskeletal disorders recruited from the Hospital for Rheumatic Diseases (Bad Wurzach, Germany), internal consistency (Cronbach's alpha) of the SIP136 was very low for sleep and rest (0.28), eating (0.33), communication (0.41), and emotional behavior (0.59); marginal for home management (0.66), work (0.64), recreation and hobbies (0.67), and mobility (0.69); and was within the acceptable range for social interaction (0.71), ambulation (0.76), alertness behavior (0.76), and body care and movement (0.80) (139). Internal consistency of the overall score was 0.83.

Test–retest information on the SIP136 in rheumatology and more generally is scant, but indicates good temporal stability of dimension and overall scores. In a sample of 49 individuals with musculoskeletal disorders who completed a second SIP 3 weeks after the initial administration, intraclass correlation coefficient (ICC) was 0.94 for the physical function dimension and 0.93 for the overall SIP136 score, indicating excellent reliability; no information was provided about test–retest reliability of the remaining scores (111). The SIP overall score also showed good temporal stability in another study, involving 130 patients with chronic low back pain. The ICC for the SIP136 total score was 0.70 over a 2-week test–retest interval (140).

Temporal stability of subscale scores seemingly had been assessed only using Pearson's correlation coefficient, with results indicating below optimal reliability for some of the scale. Correlation coefficients ranged between 0.49 (eating) and 0.86 (mobility) over a 4-week interval in a study of 299 musculoskeletal patients (139) and between 0.62 (household management) and 0.85 (ambulation) in a study involving 119 individuals with a range of chronic conditions (141). However, Pearson's correlation coefficient is unable to capture any systematic changes in scores over time, so that the actual stability of SIP136 subscale scores might be even lower.

Internal consistency reliability of the SIP68 in rheumatic conditions is not well studied, with the available evidence indicating suboptimal internal consistency, at least for some subscale scores. Internal consistency estimates (Cronbach's alpha) for the SIP68 in a study of 51 outpatients with rheumatic conditions (118) ranged from 0.49–0.87 (coefficients for specific domains not specified). However, this sample size was relatively small and generalizability of this finding is not clear. More broadly in the field of disability, for a study conducted with 111 independently living wheelchair users, Cronbach's alpha of the total SIP68 score was $\alpha = 0.88$, with scores for the individual scales ranging from $\alpha = 0.53$ (for mobility control) to $\alpha = 0.85$ (somatic autonomy) (107). However, given the much higher levels of physical disability in this sample than would be expected in rheumatic diseases, it is not known whether these results can be extrapolated to rheumatology.

Test–retest reliability of the SIP68 was assessed using a 48 hour test–retest interval in a study of 51 outpatients with rheumatic health problems using self-completed questionnaires (118). The ICCs for different categories ranged from 0.90 (mobility range) to 0.97 (somatic autonomy) and was 0.97 for the overall SIP68 score, indicating excellent test–retest reliability of this questionnaire. In another study, involving 401 people with disabilities (including arthritis), the ICCs for test–retest reliability of SIP68 were above 0.75 for all subscales and dimensions, except the physical dimension (0.61) score (142).

Validity. The items of SIP instruments appear to have good face validity as a health measure, reflecting aspects of everyday life that are likely to be affected by illness. However, no specific a priori conceptual model was used in the SIP construction, which makes it difficult to comment on its content validity. Furthermore, content validity of an instrument varies from context to context, depending on the nature of the concept being studied. Nonetheless, the SIP covers a wide range of health-related behaviors, many of which are likely to be relevant in rheumatology. Development of the SIP136 involved both patient and expert consultation, during both the item construction and selection phases of development; hence, the questionnaire is likely to be relevant to patients and clinicians. However, content validity of the SIP instruments is undermined by the presence of ceiling effects, which indicates that these questionnaires do not adequately capture the full range of health problems at a less severe end of the ill health continuum. This could potentially pose problems when assessing interventions in populations that are not severely affected by illness.

Although results for the construct validity of SIP instruments are generally favorable, given the unsatisfactory reliability of some subscales, findings about their construct validity should be viewed with caution. Factorial validity of the SIP68 and SIP136 is not well supported, with different pattern of factor loadings to that reported in the original publications generally emerging (138,142). Construct validity of the SIP136 in rheumatic conditions was supported by the expected pattern of correlations with the Arthritis Impact Measurement Scales (AIMS), a multidimensional questionnaire designed to measure ill health in arthritis. Over 12 months of followup in a study of 115 patients with knee or hip osteoarthritis (143), SIP subscales had moderate to strong correlations ($r = 0.37–0.76$) with the corresponding subscale of the AIMS ($P < 0.001$). Correlations for the total scores of the SIP136 and AIMS ranged between 0.70 and 0.73 (143). Further support for the construct validity of the SIP was found in another study, involving 299 patients with musculoskeletal conditions, where the hypothesized pattern of correlations was found between the SIP total score and a range of functional and psychosocial measures, including the Measurement of Patient Outcome Scale (arthritis-specific questionnaire in

German) (r = 0.72) and the Keitel Index of Functional Status (r = 0.60) (139).

The SIP68 also received support for its construct validity, with a correlation coefficient of 0.94 for its total score with the SIP136 in a study of 401 people with mobility disabilities, including spinal cord injury, multiple sclerosis, and arthritis (142). In another study with people with physical disabilities (n = 398), physical and psychosocial dimensions of the SIP68 had high correlations with corresponding dimensions of SIP136 (r = 0.91 and r = 0.92, respectively) (138). In the same study, construct validity of the SIP68 was further supported by the expected pattern of correlations with other generic health measures, including the SF-36 and the Katz Index of Activities of Daily Living.

Known-groups validity of the SIP instruments in rheumatology appears to be well supported. In one study (n = 172) (139), the SIP136 overall score was able to differentiate people with musculoskeletal disorders from healthy controls, although the difference was much larger for women (standardized effect size [ES] 1.07) than for men (ES 0.66). Across individual dimensions, work, eating, mobility, alertness behavior, sleep and rest, and communication were unable to differentiate between patients and controls for men, with no significant differences in the scores of these groups ($P > 0.05$). On the remaining subscales, male patients scored higher than controls, with ES ranging from 0.21 (communication) to 1.07 (body care). For women, only work was unable to differentiate between patients and controls ($P > 0.05$), with ES ranging from 0.42 (mobility) to 1.53 (home management). Deyo et al (144) have also published mean scores and SDs for the SIP overall and the physical and psychosocial dimensions that correspond with functional impairment levels of the American Rheumatism Association Functional Classification.

The SIP68 was also able to differentiate between people with spinal cord injury and those who had rheumatic diseases, with significantly worse health status scores on emotional stability, social behavior, mobility range, and psychic autonomy and communication for the rheumatic conditions groups (Z scores 2.10, 5.10, 5.71, and 2.01, respectively). Somatic autonomy and mobility control scores did not differ significantly between the study groups (145). Ability of the SIP68 to differentiate between spinal cord injury and rheumatic conditions was comparable to that of the Nottingham Health Profile.

Ability to detect change. Sensitivity of SIP instruments in rheumatology has not been well studied. Although little is known about the sensitivity of the SIP68, results indicate that the SIP136 has high specificity to detect change in health status. In a study of 79 patients with rheumatoid arthritis, the SIP136 total score had high specificity to detect a 3-point change in self-rated function, with a specificity of 0.90 for detecting worsening and a specificity of 0.76 for detecting improvement (146). However, sensitivity to improvement was 0.25, sensitivity to worsening was 0.29, the predictive score of improvement was 0.50, and the predictive score of worsening was 0.31. This indicates that a cutoff threshold of 3 points on the SIP score change had only moderate accuracy in identifying self-perceived change in health functioning.

More favorable results were obtained in a study of 54 patients undergoing joint replacement surgery, where overall and psychosocial dimension scores of SIP136 were able to detect large improvement at 6-months postsurgery, with standardized response means (SRMs) of 0.94 and 0.88 (147). As might be expected with surgical intervention, smaller improvement (SRM 0.77) was recorded for psychosocial dimension. Similar results were obtained in another orthopedic surgery cohort at 1-year followup (148). SIP136 overall and physical dimension scores were also able to detect self-reported change in health status in a sample of 127 musculoskeletal patients, with an ES of 0.42 and 0.39, respectively (111). In another study involving 299 musculoskeletal patients, SIP overall, body care and movement, emotional behavior, and sleep and rest scores showed small improvements (ES 0.20–0.28) following 4 weeks of conservative treatment (139). Statistically significant improvements were also found on alertness behavior, ambulation, home management, social interaction, and mobility subscales, although these changes failed to reach practical importance (ES <0.20); communication, recreation and hobbies, eating, and work subscales showed no change over the study period.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SIP includes several items relevant in rheumatology settings including mobility, pain, and functional capacity. As generic measures of health status, SIP instruments would be useful when comparisons of the impact of rheumatic disease on individual health status with that of other illness is required. There is also good evidence to suggest that the SIP (especially the SIP136) is able to detect change in a range of interventions in rheumatology.

Caveats and cautions. Several studies have revealed considerable weaknesses of the SIP68 and SIP136, particularly in the area of reliability. Both versions of the SIP also exhibit severe ceiling effects, which suggest that these instruments may not be useful for low to moderate levels of health impairment.

Clinical/research usability. The low reliability of some subscales in the SIP indicates that the overall score, rather than subscale scores, is more likely to return more robust data. Administration and responder burden may be barriers to clinical use of the SIP136; however, this appears to have been rectified in the SIP68. The relatively high cost of the SIP136 may further limit its usability in clinical settings and research settings. Given the comparable psychometric properties of the 2 versions of the SIP and the greater administrative burden and cost of the SIP136, it appears there are no advantages of using the SIP136 over the SIP68.

INTRODUCTION

Health Utility Measures

The EuroQol 5-domain, Short Form 6D, Health Utility Index Mark 3, the Quality of Wellbeing Scale, and the

Assessment of Quality of Life Scale are health utility measures of generic health-related quality of life (HRQOL) originating from the field of health economics. These scales are also defined as multi-attribute utility instruments, which means that they consider multiple independent attributes of an individual to create an indication of overall HRQOL, ranging from perfect health (1.0) to death (0.0), and may even include states worse than death (<0.0). The individual attributes of HRQOL contained within the questionnaire (often described as the "descriptive system") are weighted by society's strength of preference for those health states. The strength of preference is termed the utility of a health state and is obtained by asking members of the community to rank desirability of a given health state relative to perfect health and death. The utility of health states represented in the descriptive systems is generally achieved through specialized interviews such as time trade-off or standard gamble.

Although all HRQOL instruments purport to measure the same thing, across the perfect health to death continuum, they often do not. The values obtained from each for the same health state (person- or community-tested) vary because each instrument has different content, and different weights are used to generate its overall utility score (149–153). Since each instrument can generate a different value, different change scores will be obtained across instruments (154–157). Given this, the choice of instrument included in a study has the potential to generate results suggesting a null or positive result (158–160), although Ruchlin et al suggest that there is no specific pattern emerging (161). The recent work of Seymour et al suggest that choosing an instrument is difficult without good prior information surrounding the expected magnitude and direction of health improvement related to a health care intervention (162).

MEDICAL OUTCOMES STUDY SHORT FORM 6D (SF-6D)

Description

Purpose. The SF-6D utility score is derived from items within the widely used SF-36 and SF-12. The purpose of the SF-6D is to provide ratings of an individual's health-related quality of life (HRQOL) across all health conditions. The ratings of HRQOL are also called "utilities" or preferences for health states that are used in health economic evaluation and to derive quality-adjusted life years (QALYs) for use in cost utility analysis. Brazier et al published an algorithm for estimating SF-6D utilities scores from the SF-36 in 2002 (163) and from the SF-12 in 2004 (164). The initial scoring algorithms were updated in 2008 (www.shef.ac.uk/scharr/sections/heds/mvh/sf-6d).

Content. The SF-6D covers 6 domains, including physical function, role limitation, social function, bodily pain, mental health, and vitality.

Number of items. The SF-6D utility score can be derived from 11 items of the SF-36 or from 7 items of the SF-12 (164).

Response options/scale. Items are scored on a Guttman scale, where the health states have increasing severity (disutility) expressed as limitations (in activities, the kind of work one can do, social activities), degree of pain interference with daily life, frequency of feeling down-hearted, or frequency of feeling fatigued. The number of response levels on the SF-6D items ranges between 3 and 5 (SF-12 derivation [164]) or 4 and 6 levels (SF-36 derivation [163]).

Recall period for items. The SF-6D is available in 4-week or 1-week recall periods.

Endorsements. The utilities derived from the SF-6D are being used in a wide range of health economic studies to provide estimates of cost per QALY, therefore enabling comparison of alternative treatments. These data inform policy makers of the relative value of new interventions. In several countries including the UK, Australia, and New Zealand, the calculation of QALYs is essential for economic evaluations of pharmaceuticals submitted to the government agencies (i.e., National Institute for Health and Clinical Excellence in the UK).

Examples of use. Several studies have used the SF-6D to estimate QALYs, therefore providing the economic dimension in treatment effectiveness studies, including studies of tumor necrosis factor–blocking agents (12,165), spa treatment for people with fibromyalgia (166), a physical exercise program for people with rheumatoid arthritis (RA) (167), and in a survey to express the burden of disease of people with RA (168).

Practical Application

How to obtain. The SF-6D can be obtained from SF-36 or SF-12 questionnaire scores, therefore it is necessary to obtain these questionnaires. For details, see the How To Obtain section for the SF-36 and SF-12 in this issue.

Method of administration. As with the SF-36 or SF-12, the SF-6D can be self- or interviewer-administered. For details, see the Method of Administration section for the SF-36 and SF-12 in this issue.

Scoring. The SF-6D utility score is calculated as a function of weighted scores across the items that comprise this tool. The algorithm to obtain SF-6D scores from the SF-36 and SF-12 questionnaire data can be obtained through 3 types of licenses, as described on the University of Sheffield web site (www.sheffield.ac.uk/scharr/sections/heds/mvh/sf-6d): 1) a license is available free of charge for all noncommercial applications including work funded by research councils, government agencies and charities, 2) for commercial applications there is a per-study license (e.g., clinical trial), although an open license for a fixed period is available, and 3) the SF-6D can be calculated using purpose-developed software available from QualityMetric.

Score interpretation. The SF-6D produces an interval scale utility score, ranging from 0.30 (poor HRQOL) to 1.0 (perfect health). The SF-6D utility measure can also be used as an indicator of relative disease burden across diseases. This is dependent on reliable population norms being available, such as those proposed by Fryback et al for the US (169). Uhlig and colleagues used the SF-6D to compare the HRQOL of people on the Oslo Rheumatoid Arthritis Register with people from the general population and found that people with RA have 0.16 lower utility than the population. They were therefore able to present a case that RA contributes a substantial disease burden on individuals and society (168).

Respondent burden. See the Respondent Burden section for the SF-36 and SF-12 in this issue.

Administrative burden. See the Administrative Burden section for the SF-36 and SF-12 in this issue.

Translations/adaptations. The SF-36 is available in 121 languages, therefore the SF-6D is similarly available. Specific information can be obtained from the International Quality of Life Assessment web site, http://www.iqola.org.

Psychometric Information

Method of development. The SF-36 and SF-12 items were revised to only cover 6 dimensions of health while maintaining maximum coverage of the original breadth of the questionnaires. The challenge for the developers was to provide valuations of all the different combination of health states that could be represented across the 6 items, each with 3 or more levels. The total number of possible health state combinations is 18,000, which is far too many to value in practice. A common procedure in health economics is to select a minimum range of these using an orthogonal design, and therefore infer the valuations of the health states not directly valued. A total of 49 combinations of levels of the 6 items was valued by a representative community sample using a technique called standard gamble (SG) (170). In the SG interview, the respondent is asked to choose between the certain prospect (A) of living in an intermediate state defined by the SF-6D and the uncertain prospect (B) of 2 possible outcomes, the best state defined by the SF-6D or the worst state. The chances of the best outcome occurring is varied until the respondent is indifferent between the certain and uncertain prospects. The data obtained from these valuations are then used in various modeling procedures to generate an algorithm to convert the SF questionnaires into SF-6D utility scores. Further details are available from the development papers (163,164).

Acceptability. The acceptability and missing values of the SF-6D are reflected in the original questionnaires (the SF-36 and SF-12), which are generally acceptable. Barton and colleagues compared the completion rates of the SF-6D with the EuroQol 5-domain (EQ-5D) measure in 1,865 general practice patients and found that individuals who were older, women, of a lower occupational skill level, from an area of lower socioeconomic status, or used prescribed medication were significantly less likely to complete the SF-6D (84%) compared with the EQ-5D (93%) (171).

Importantly, HRQOL measures are intended to provide valuations of health states that range from perfect health (1.0) to death (0.0). However, the SF-6D scale does not extend beyond 0.3, i.e., the worse health state described by the SF-6D does not extend to death. This is a serious flaw if a substantial number of subjects in a study are expected to have very poor health states.

Reliability. The reliability of the SF-6D has been tested in a variety of settings, with generally favorable results. In a small study (n = 61) of proximal humeral fractures, Slobogean et al (172) found good reliability (intraclass correlation coefficient [ICC]) for the SF-6D (0.79) and EQ-5D (0.78), but poor for the Health Utilities Index Mark 3 (0.47). Khanna et al also found the SF-6D to be reliable in a sample of patients with systemic sclerosis (ICC 0.82) (173). On the other hand, Boonen et al found that in patients with ankylosing spondylitis, the test–retest reliability of the SF-6D was only modest (ICC 0.68), and this was greatly reduced in subgroups with lower disease activity (174).

Validity. The SF-6D contains items that cover physical function, role limitation, social function, bodily pain, mental health, and vitality. Consistent with most utility scales, the SF-6D was not derived through consultations with patients and clinicians to ensure face and content validity (151). Nonetheless, the dimensions are broadly concurrent with those covered by the many disease-specific tools available in rheumatology. Support for convergent and discriminant validity of the SF-6D is evidenced by consistent findings of moderate correlations between the SD-6D and other HRQOL scales (175–179) and lower, but still substantive, correlations with disease-specific questionnaires (174,177,179,180).

The known-groups validity of the SF-6D appears to be supported. Marra et al undertook a comprehensive study in 313 people with RA to compare several disease-specific measures (Rheumatoid Arthritis Quality of Life Questionnaire and the Health Assessment Questionnaire [HAQ]) with several preference-based measures including the SF-6D (179). They found that utility scales, including the SF-6D, appeared to discriminate well across RA severity categories, although the disease-specific measures were generally more sensitive in this setting. In 167 patients with systemic lupus erythematosus, Aggarwal et al (178) found that both the EQ-5D and SF-6D tools differentiated among patient groups of varied disease severity. Importantly, very few patients in this study reported very low HRQOL, therefore the tools are more likely to appear to perform relatively well. However, in a population sample, the SF-6D has been found to be more sensitive than the EQ-5D in detecting differences between groups of individuals reporting very good, good, fair, bad, or very bad health (181).

Ability to detect change. Evidence for the ability of the SF-6D to detect change is mixed. While several studies have demonstrated that the SF-6D is capable of detecting change, findings of many other studies are less favorable. Boonen et al found that in 254 patients with ankylosing spondylitis, the smallest detectable change was smaller (i.e., more sensitive) in the SF-6D compared with the EQ-5D. However, it discriminated less well between patients with different disease severities (174). Harrison et al undertook a comparative responsiveness study of the EQ-5D and SF-6D in cohorts of patients with early inflammatory disease through to severe RA (182). As the use of the SF-6D in patients with severe progressive disease may be inappropriate due to the scale not extending lower than a utility of 0.30, the study by Harrison and colleagues (182) highlights the need for careful attention to disease severity at study onset. The SF-6D did, however, appear to be somewhat more responsive than the EQ-5D in detecting improvements in health (182). On the other hand, in a controlled trial, Barton and colleagues administered the Western Ontario and McMaster Universities Osteoarthritis Index to 389 people with knee pain and classified change score as no change, improved >20%, or declined >20% (158). The SF-6D performed poorly at detecting improvement. Similar results were obtained by Adams et al in 505 patients with RA and psoriatic arthritis and again reflect the inability of the SF-6D to detect poor health states (183).

Several studies reported on minimum clinically important difference (MCID) of the SF-6D. In rheumatology settings, Khanna et al have proposed a MCID of 0.035 units in systemic sclerosis using change in the HAQ Disability Index score as an anchor (173), and Marra et al have estimated a MCID of 0.03 for people with RA using the SF-36 health transition question as an anchor (179). More broadly, Walters and Brazier undertook a review of 11 studies across a variety of health conditions and found that the MCID for the SF-6D ranged from 0.011–0.097, with a mean of 0.041. The corresponding standardized response means ranged from 0.12–0.87, with a mean of 0.39, and were in the "small to moderate" range using Cohen's criteria (152).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SF-6D can be a useful indicator of utility in the absence of other utility measures. A unique aspect of this tool is that, if the SF-36 and the SF-12 have been applied in a completed trial or observational study, a utility score for cost utility analyses can be derived from existing data without the need for administering further questionnaires.

Caveats and cautions. The major drawback of the SF-6D is that the scale does not cover the range from below 0.3, which would be a common health state in many rheumatic conditions. This makes the scale insensitive to changes between very poor health and moderate health. If researchers are working with a well-defined clinical condition with mild to moderately poor HRQOL, then the SF-6D may be preferred over other utility measures, such as the EQ-5D, which may be insensitive to improvements in this range.

Clinical/research usability. The SF-6D is not a tool to be used in the clinical setting since it is a utility instrument designed to inform economic evaluations. It is also useful for comparisons across conditions, and to provide estimates of relative societal burden of different conditions when national norms are used as benchmarks.

HEALTH UTILITIES INDEX MARK 3 (HUI3)

Description

Purpose. The HUI is a family of generic preference-based (utility) measures developed for measuring health-related quality of life (HRQOL) (184). The intended uses of the HUI include describing treatment processes and outcomes in clinical studies, economic evaluations of health care programs, and the measurement and monitoring of population health (185). The original version (HUI1) was developed for assessment of out-of-pocket costs and quality of life of pediatric oncology survivors. The HUI2 was developed as a revised version of HUI1 to measure the global morbidity burden of childhood cancer (184). The HUI3 was developed as a more generically applicable measure than HUI1 and HUI2. Items present in earlier versions specific to pediatrics (e.g., cognition domain items in HUI2 relate to schoolwork) were replaced with more broadly applicable items, while some domains were expanded (e.g., sensation in HUI2 was broken into 3 separate domains of vision, hearing, and speech in HUI3), and others were removed (e.g., fertility in HUI2 was removed from HUI3). Therefore, HUI3 domains largely overlap with those of HUI2. HUI1 was first published in 1982 (186), while HUI2 and HUI3 were described in the literature in the mid-1990s (187). This review focuses on HUI3, since this version is commonly used in rheumatology (117,188).

Content. The HUI3 measures 8 HRQOL domain areas including vision, hearing, speech, ambulation/mobility, pain, dexterity, emotion, and cognition. HUI2 measures the 7 domain areas of sensation, mobility, emotion, cognition, self-care, pain, and fertility.

Number of items. HUI2 and 3 require the participant to select one descriptor that most accurately reflects their condition per domain.

Response options/scale. Each domain within the HUI3 has 5–6 rank-ordered response options, while HUI2 has 3–5 response categories per domain. Descriptors of response categories may contain 1 element (e.g., the HUI3 emotion domain has a response option of "somewhat happy") or it may contain several elements (e.g., the HUI3 hearing domain has a response option of "able to hear what is said in a conversation with one other person in a quiet room with a hearing aid," and "able to hear what is said in a group conversation with at least three other people, with a hearing aid"). Therefore, if the HUI3 is being administered via telephone where the participant cannot read the entire descriptor of a response option, a series of shorter questions need to be asked to allow the individual to select a response option that is most appropriate to their situation.

To resolve this problem, the developers have produced a 15-question (15Q) survey to allow the participant to identify the appropriate response option based on a series of shorter questions. There is also a 40-question (40Q) survey comprised of even less complex, predominantly yes/no response options. The 40Q version of the HUI3 has a skip pattern so that only some questions will need to be asked of each participant.

Recall period for items. There are several versions of the HUI3 available with recall periods of 1 week, 2 weeks or 4 weeks (e.g., "Describe your ability during the past 4 weeks to . . ."). There is also a version available for "usual health," where participants are asked about their usual health (e.g., "Describe your usual ability to . . .").

Examples of use. In rheumatology, the HUI3 has been used to assess HRQOL in patients with rheumatoid arthritis (188,189), changes in HRQOL in patients with juvenile idiopathic arthritis (190), and to assess the effectiveness of hylan G-F 20 in treatment of knee osteoarthritis (191).

Practical Application

How to obtain. The HUI2 and HUI3 classification systems can be viewed online at http://www.healthutilities.com (185). The questionnaires and user manuals are only distributed under license from HUInc. The cost of HUI3 is $4,000 per study for the questionnaire and the matching user manual (185).

Method of administration. The 15Q version of the HUI3 is designed to be self-administered, while the 40Q version can be interviewer-administered (by telephone or face-to-face), on paper or using a computer (184).

Scoring. The functions to derive the scores are multiplicative and based on classical utility theory. The scoring manual contains decision tables showing all possible combinations of responses per attribute. Typically, scoring is done using a common statistical package such as SPSS or SAS. A spreadsheet such as Excel can be used, but it is not recommended by HUI developers if there are more than a few subjects and/or multiple assessment points. The decision tables of response combinations are used to determine the health-state level for each health domain and then, using the tables and the scoring algorithm, the utility scores for all attributes of health and the overall HRQOL score can be determined.

Missing responses are scored as 0. At the same time, the presence of missing responses is problematic, since at least 2 scores (1 domain and the overall score) will be missing for each subject that has 1 response missing. Nonresponse to an item on the 40Q will also cause problems with the skip pattern, making the questionnaire difficult to score.

Score interpretation. The score range for HUI3 is −0.36–1.00 and −0.03–1.00 for HUI2. A score of 1.00 signifies perfect health and 0.00 represents death. HUI allows for negative numbers for health states considered worse than death. Population normative data are available from numerous large general population surveys. Normative values by age (15+, 17+, 18+, 20−85 and 35–89 years), race ("unselected," "Hispanic non-Black," "Black non-Hispanic," "non-Black and non-Hispanic") and country (Canada, USA) are available on the HUI web site (185).

Respondent burden. HUI3 generally has low responder burden. The mean time to complete the 15Q is 5–10 minutes (151). The 40Q, which has a built-in skip pattern takes 3 minutes to complete (184).

Administrative burden. Administration burden for the HUI3 is moderately high. Interviewer administered assessments will require interviewer training, especially for the 40Q version of the HUI3. It is also recommended to review completed 15Q version questionnaires once received and to contact the respondent if there are missing answers. Scoring will require basic knowledge of statistical software.

Translations/adaptations. HUI3 was first developed in English and is now available in more than 35 languages worldwide. It has been used successfully without modification in Canada, the UK, the US, and Australia. There are 16 variations of the HUI questionnaires, which are dependent on mode of administration (self-complete or interviewer-administered), recall period (past week, 2 weeks, 4 weeks, or usual health), and assessment viewpoint (self or proxy). One or more variations of the HUI questionnaires are available in Afrikaans, Chinese (traditional and simplified characters), Croatian, Czech, Danish, Dutch, Finnish, Flemish, French (continental or European French and French-Canadian), German, Hebrew, Hungarian, Italian, Japanese, Korean, Malay, Norwegian, Polish, Portuguese (European and Brazilian), Romanian, Russian, Serbian, Slovak, Spanish (European and Mexico, Latin and South American), Swedish, Thai, and Turkish. Other versions in preparation include Serbian.

Psychometric Information

Method of development. The original items for the HUI were generated from the work of Cadman et al (192) (note, we have been unable to access this original work from 1986), who sought to determine the most important attributes of HRQOL based upon clinical experience. A random sample of adults from the general population then ranked these attributes on their desirability (193); this information was subsequently used to derive weights for the HUI2 and HUI3.

Acceptability. Given the complexity of some of the response option descriptors within domains of the HUI3, the 15Q and 40Q (with simplified response options) have been developed to make it easier for participants (or the interviewer administering the HUI3) to select an appropriate response option descriptor. Missing data on HUI3 in rheumatology studies are not frequent. A study among 114 rheumatology outpatients found that there were no missing responses at a baseline face-to-face assessment on HUI3 administered by a trained nurse interviewer. In a telephone-based followup interview 2 weeks later, <5% of the respondents had missing data (194). Similarly, floor and ceiling effects are not commonly encountered in rheumatology populations. Only 4 subjects (3.5%) in the above study obtained the highest possible health rating.

Reliability. Results for reliability of the HUI vary considerably. Cronbach's alpha ($\alpha = 0.71–0.79$) was reported for the Spanish version of the HUI3 in the general population (195). For a cohort of heart-failure patients, Cronbach's alpha for the total score of the HUI3 was reported as $\alpha = 0.51$ (196). We have not been able to identify any studies that assessed internal consistency of the HUI3 in rheumatology-specific populations. Test–retest reliability of interviewer-administered HUI3 in a study of 114 rheumatology outpatients was intraclass correlation coefficient (ICC) 0.75 (95% confidence [95% CI] 0.65, 0.83) over a 2-week period (194). However, these results are difficult to interpret, since the first interview was done face-to-face while the second interview took place over the telephone.

More favorable results were obtained in a study of 50 rheumatoid arthritis patients (randomly selected from a larger study), where 3 months test–retest reliability of HUI3 was found to be acceptable (ICC 0.81, 95% CI 0.66, 0.90) (197). Similarly, in a stratified random sample of people completing the Canadian General Social Survey (n = 506), the test–retest reliability for the HUI3 of ICC 0.77 was recorded for telephone assessments conducted 1 month apart (198).

Validity. The results of studies investigating construct validity of the HUI3 are mixed. An observational study of 144 rheumatology outpatients (194) found that HUI3 did not discriminate between people with and without chronic health conditions. Despite the hypothesized high to moderate correlations, the correlations between HUI3 and Short Form 36 (SF-36) scores were in a low to moderate range ($\rho = 0.29-0.49$, $P < 0.01$ for all), with the SF-36 physical functioning and bodily pain scales showing the lowest and highest correlations with the HUI3 score, respectively. When compared to the EuroQoL 5-domain (EQ-5D) instrument, median EQ-5D and HUI3 scores were very similar. The correlation between EQ-5D and HUI3 scores for all patients was $\rho = 0.45$ for baseline interviews and $\rho = 0.57$ for followup interviews (Spearman's rho, $P < 0.001$ for both).

On the other hand, a study of 114 osteoarthritis patients on the waiting list to see an orthopedic surgeon (199) found support for construct (convergent and discriminant) validity of both HUI2 and HUI3. Of the 87 a priori hypotheses examined, 75% were confirmed by zero-order correlations, suggesting that the constructs within the HUI2 and HUI3 were, in general, related to similar constructs in other conceptually related measures (SF-36, Harris Hip Scale [HHS], Western Ontario and McMaster Universities Osteoarthritis Index [WOMAC], McMaster Toronto Arthritis Patient Preference Questionnaire, the State-Trait Anxiety Inventory, and the 6-Minute Walk Test).

The HUI3 was used in the 1990 Ontario Health Survey and found to be able to differentiate people with stroke or arthritis from those who had neither of these conditions (200). The highest mean score, indicative of best health, was for people without a history of arthritis or stroke (0.93), followed by those who had arthritis (0.77) and stroke (0.54).

Ability to detect change. Results for the ability of the HUI3 to detect change are also mixed. In a study of 99 patients on a waiting list for total hip arthroplasty who had completed the HUI3 before and after the surgery (201), the HUI3 showed improvement in the overall summary score and various domains following surgery. There was a large standardized effect size (ES) for the overall summary score (1.19) and pain (1.30), and a moderate ES for ambulation (0.56). There was no change in vision, hearing, speech, dexterity, and cognition, which would be expected in this population. Although the HUI3 was not as responsive to change after total hip arthroplasty as the disease specific measures considered in the same study, (HHS, WOMAC), it was the most responsive of the generic measures considered (SF-36, EQ-5D, and HUI2).

Less favorable results for the responsiveness of the HUI3 were obtained in a study of 320 rheumatoid arthritis patients recruited from private rheumatology practices (197). The study compared responsiveness to change over time (disease progression) of a number of generic HRQOL measures (HUI2, HUI3, SF-6D, EQ-5D) as well as some disease-specific measures (the Health Assessment Questionnaire Disability Index and the Rheumatoid Arthritis Quality of Life Questionnaire). The HUI3 appeared to be poorly responsive to deterioration but was able to identify those classified as "better" on global assessment of disease severity at 3- and 6-months followup. Of all the measures used, the HUI3 and SF-6D were found to be the most responsive between baseline and 6 months for measuring improvement ("worse" HUI3 = ES −0.10, 95% CI −0.31, 0.13; "same" HUI3 = ES 0.12, 95% CI −0.03, 0.26; "better" HUI3 = ES 0.23, 95% CI 0.08, 0.41) (197).

Information on minimum clinically important difference (MCID) for HUI3 in rheumatology is limited. In a study with individuals who had stroke or arthritis, drawn from the 1990 Ontario Health Survey, MCID on the HUI3 was defined as a difference of 1 level within HUI3 attributes, which equates to a change of ≥0.03 units in the HUI3 score. More generally, Drummond reported that a difference of ≥0.03 in mean HUI overall HRQOL scores were clinically important, and differences as little as 0.01 may be meaningful and important in some contexts (202). However, it is not clear how these values were derived.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument appears to measure some aspects of quality of life that are affected by rheumatic diseases, although there are several items (e.g., hearing) within the scale that are not relevant to this field. As a utility measure, HUI3 can be used in health economic analyses. The instrument appears to be sensitive to positive changes brought about by some treatments for rheumatic conditions (e.g., hip replacement). However, it appears to be poorly responsive to deterioration, and therefore may not be suitable for individual followup. This instrument appears to be widely applicable to most patient populations; however, research to date in rheumatology has been primarily in rheumatoid arthritis and total hip replacement populations.

Caveats and cautions. This instrument does not appear to be as sensitive to change brought about by treatment of disease as other disease- or joint-specific instruments. There may be difficulties using this instrument among older adult populations or persons with cognitive impairment due to the complexity of some of the items. The psychometric robustness of the HUI3, especially its temporal stability and construct validity, have also received mixed support in rheumatology.

Clinical/research usability. The interpretation of HUI3 scores in clinical settings is hampered by the lack of information on cutoff scores for what is considered to be meaningful change in HRQOL for patients with rheumatic conditions. For example, there is no cutoff threshold indicative of when joint replacement may be required or whether such surgery has been successful at improving an individual's HRQOL. The HUI appears to have a moderate administrative burden, although use of the computerized scoring algorithm may compensate for the extra interviewer training necessary for the administration of the 40Q version. Respondent burden does not appear to be a problem that would limit clinical or research use. The cost of the questionnaire and scoring algorithms may limit the use of the HUI3 for clinician-initiated unfunded research projects.

QUALITY OF WELL-BEING SCALE (QWB)

Description

Purpose. The QWB scale was developed more than 30 years ago as a measure of health-related quality of life (HRQOL) in the general population (203). The QWB is a preference-based measure that combines functioning and symptoms to produce a well-being index ranging from 0 (death) to 1 (full, symptom-free functioning) (204). The QWB can also be used to calculate quality-adjusted life years, which combine life expectancy with HRQOL to produce a summary measure of quality and quantity of life lived.

The QWB was initially developed for interviewer administration, but the use of this measurement tool has been low due to its length and difficulty of administration (205). The self-administered version, Quality of Well-Being Scale-Self Administered (QWB-SA) was developed to address these limitations of the QWB. The QWB-SA was released in 1997 (204). This review focuses on the QWB-SA version of the instrument.

Content. The QWB-SA includes a wide range of physical and mental symptoms that people might experience in daily life (205). The symptoms assessed by QWB-SA reflect different aspects of health and cover different degrees of severity. Most items focus on a specific problem related to one body system, such as visual problems (e.g., blindness) or central nervous system functioning (e.g., paralysis).

The QWB-SA has 5 parts, including a symptoms checklist and 4 function sections. The symptoms section incorporates assessment of chronic (e.g., speech problems, physical deformities) and acute symptoms. Acute symptoms include physical (e.g., headache, pain) and mental health symptoms (e.g., sadness, anxiety). The function sections of the QWB-SA include self-care, mobility (including use of transportation), physical activity (e.g., climbing stairs), and usual activity (e.g., work, home, or recreation) (205).

Number of items. The QWB-SA consists of 74 items. The symptom checklist has 58 symptoms, including 19 chronic symptoms, 25 acute physical symptoms, and 14 mental health symptoms. Self-care is assessed by 2 items, the mobility and usual activity sections have 3 items each, and the physical activity section has 8 items.

Response options/scale. The presence/absence of 19 chronic symptoms is measured on a dichotomous scale (yes/no), with participants asked to indicate whether they are currently experiencing any of the symptoms or problems listed. For the remaining items, participants are asked to indicate which days over the past 3 days they experienced each of the health problems listed, using a 4-point scale with response options including "no days," "yesterday," "2 days ago," and "3 days ago." Respondents are able to select more than one response option if they experienced the symptom on more than one of the days (for example, yesterday and 3 days ago). Responses are scored according to the number of days that a health problem was experienced (0, 1, 2, or 3).

Recall period for items. With the exception of the chronic symptoms section, QWB-SA asks patients about symptoms and function over 3 days prior to the day of administration. The format of the chronic symptoms questions does not use the 3-days recall period since it is expected that chronic conditions do not vary much over the 3-day assessment period (205).

Endorsements. Approved by the Scientific Advisory Committee of the Medical Outcomes Trust (http://www.outcomes-trust.org/instruments.htm).

Examples of use. In rheumatology, the QWB/QWB-SA had been previously used to measure HRQOL in osteoarthritis (206), to measure the impact of total hip or knee replacement on HRQOL (81,207), and to assess the impact of an active drug treatment relative to placebo on HRQOL in a randomized controlled trial in rheumatoid arthritis (208).

Practical Application

How to obtain. An inspection copy of the QWB-SA can be obtained from https://hoap.ucsd.edu/qwb-info/. For nonprofit organizations, the scale and scoring instructions are available free of charge, although the researchers are required to sign a copyright agreement with the Health Services Research Centre (HSCR), the University of California, San Diego. Profit organizations are required to purchase a yearly license at $1,000 per year, with an additional charge of $0.25 for each questionnaire administered.

Method of administration. The QWB-SA was designed for self-administration and is available in paper and pencil or web-based formats. The QWB-SA can also be administered by telephone or in a face-to-face interview, although the psychometric properties of these methods of administration have not been specifically studied (205).

Scoring. The QWB-SA requires computerized scoring. A scoring algorithm (SPSS syntax) is available for purchase from the scale developers for $240. The QWB-SA scoring algorithm assumes that missing responses are equivalent to the absence of a problem.

Score interpretation. Symptoms and the 4 function scores are combined into a total preference-weighted score of well-being that ranges from 0 (death) to 1.0 (symptom-free, optimal functioning). Normative data are available for clinical and nonclinical samples by age, sex, and ethnicity. However, these normative data, especially for nonclinical samples, are based on relatively small numbers of participants, with a total normative sample of 843 people. Participant numbers across subgroups range from 1 (e.g., Native Americans age ≤30 years) to 235 (whites age ≥71 years) (205). Another recent study also presents means and SEs for QWB-SA scores derived from a probability sample of 3,844 US adults ages 35–89 years by sex and 5-year age groups (169).

Respondent burden. The QWB-SA takes ~10 minutes to complete in paper and pencil format. Completion instructions are self-explanatory. In a telephone-administered interview of 3,844 US residents, completion time for the QWB-SA varied from 7.7 to 17.5 minutes with an average of 11.1 minutes (169).

Administrative burden. Administration instructions for the paper and pencil version of the QWB-SA are self-explanatory. Scoring requires access to a computer. Apart from some knowledge of SPSS statistical software, no specific training is required for administration and scoring of the QWB-SA. The QWB-SA form is designed for optical scanning, and the HSRC also provides data cleaning, as well as entry and scoring services for the QWB-SA for $57 per hour.

Translations/adaptations. The instrument is available in English, German, French, Dutch, Italian, and Spanish. Translations to other language are available upon request, with fees determined by the languages requested and project timelines/needs.

Psychometric Information

Method of development. The original, interviewer-administered version of the QWB was developed for the purpose of defining "the universe of all possible health states between optimum function and death" (203). Items for the inclusion into the QWB were generated from specialty-by-specialty review of medical reference works. The initial tool included the assessment of 3 dimensions of functioning, reflecting different levels of mobility, physical activity, and social activity, as well as 36 different health symptoms. The 3 dimensions generated 100 theoretical combinations of health states, of which 43 were observed in a pragmatic study of more than 10,000 people. Open-ended questions administered to the observational sample identified no additional health states or symptoms (203).

In the development of the QWB-SA, the symptom checklist was expanded to 58 symptoms, including at least 12 mental health symptoms (205). These additional symptoms were identified through focus groups conducted with physicians. Preference weights for the QWB-SA were derived from 435 English-speaking adults drawn from primary care clinics and college campuses in San Diego, California. Participants were presented with descriptions of hypothetical health states defined by the scale items and asked to provide numerical ratings (on a scale of 0–100) for how undesirable each health state was. These ratings were analyzed with regression analysis using levels of functioning and symptoms as predictors. Regression coefficients were subsequently used to generate weights for the scale scores (205). No specific patient groups were involved in item and weight generation for either the QWB or the QWB-SA.

Acceptability. Readability of the QWB-SA could potentially be problematic, as the scale contains words and phrases that might not be commonly understood by people with lower education levels (e.g., pelvic cramping, usual activities). The sentence structure of the QWB-SA is also rather complicated, with each item containing several concepts (e.g., "Because of any physical or emotional health reasons, on which days did you avoid or feel limited in doing some of your usual activities, such as visiting family or friends, hobbies, shopping, recreational, or religious activities."). The complicated wording and sentence structure of the QWB-SA could potentially lead to difficulties with understanding the meaning of the question, as well as difficulties with the selection of the appropriate response option, especially when used with the elderly or unwell individuals.

In a study conducted in Germany with 264 rehabilitation inpatients with musculoskeletal (n = 106), cardiovascular (n = 88), or psychosomatic disorders (n = 70), no missing data were observed for the QWB-SA (209). For comparison, the proportion of missing data on other generic HRQOL measures used in the same study was 1.3% for the EuroQol 5-domain (EQ-5D) measure and 15D, 1.9% for the Health Utilities Index Mark 3 (HUI3), and 6.1% for the Short Form 6D (SF-6D). In a random sample of the US general population (n = 3,844) (169), the proportion of missing data for the QWB-SA was 2.2%, 0.7% for the EQ-5D, 7% for the HUI3, and 2.7% for the SF-6D.

In a population-based sample of 293 adults age ≥65 years, the proportion of missing data on QWB-SA items ranged from 0.3% (hearing and skin problems items in symptoms section) to 14.6% for loss of sexual interest or performance (also a symptoms section item) (136). Nearly 50% of all respondents skipped at least 1 symptom on the QWB-SA 3-day recall section, with the mean number of missing items being unrelated to age, but higher in men than in women.

The QWB-SA appears to have good range of score functioning, with no floor or ceiling effects observed in a random sample of the US general population (n = 3,844) (169), as well as in a sample of patients from Germany with musculoskeletal conditions (n = 106) (209). While none of the HRQOL measures used in the second study showed evidence of floor effects, 5.7% of patients obtained the maximum EQ-5D score (ceiling effect), while 2 patients (1.9%) achieved the maximum possible score on the 15D and 1 patient (0.9%) on the SF-6D (209).

In a study of performance of the QWB-SA in 293 people age ≥65 years, the scale appears to have been well received by the respondents, with 60% reporting that they were very or somewhat satisfied (95% confidence interval 54.2–65.4%) with the scale. The satisfaction ratings for the QWB-SA were similar to those for the Sickness Impact Profile (SIP) (69% were very or somewhat satisfied) and the SF-36 (67% were very or somewhat satisfied) (136).

Reliability. Reliability of the QWB-SA has not been well studied in general (210), and there appears to be no published reliability data for rheumatology populations. In other clinical populations, the QWB-SA was reported to have low temporal stability, with an intraclass correlation coefficient (ICC) of only 0.59 between 1- and 6-months postoperative scores of 265 cataract surgery patients (211). However, it is possible that during the 5-months followup, real changes in the individual's HRQOL might have occurred. In an earlier study of 218 adults with stable health conditions recruited from primary care clinics, QWB-SA scores were only moderately stable over a 1-month test–retest period (Pearson's r = 0.77) (204).

Validity. The QWB-SA appears to have good face validity, with items appearing to capture health-related symptoms. Although the content validity of a measure is influenced by the nature of the construct that is being measured, the original version of the QWB was reported to have good content validity for capturing health-related symptoms. In a sample of more than 10,000 people drawn from a variety of clinical settings (203), open-ended questions (designed to elicit additional information about health-related problems that people might experience in daily life) yielded no health states or symptoms in addition to those already listed in the scale. The involvement of physicians into focus groups during QWB-SA construction (to identify aspects of health that are understood by physicians to be signs/predictors of various diseases) increased the likelihood that the scale has good content validity for use in clinical settings. The QWB-SA also appears to have good ability to capture the full range of HRQOL impairment as indicated by no or low floor/ceiling effects in the general population and musculoskeletal patients (169,209). Although the scale also contains items that are not specifically related to rheumatic conditions, retention of these items is justifiable since they represent part of HRQOL and are potentially relevant indicators of the overall well-being of people with rheumatic conditions.

Criterion-related validity of HRQOL questionnaires is difficult to establish due to the absence of a "gold standard" measure of HRQOL. Evidence for the construct validity of QWB-SA in rheumatic conditions is generally positive. While agreement between QWB-SA and other generic measures of HRQOL in musculoskeletal patients was reported to be poor to moderate (with an ICC ranging from 0.26 for agreement between QWB-SA and EQ-5D and 0.48 for agreement between QWB-SA and 15D [209]), in a community sample of older adults, QWB-SA was found to have moderate correlations with physical health components of the SIP and SF-36 ($r = \geq 0.42$) and weaker correlations with the SIP psychosocial dimension and the SF-36 summary mental health score (136). All correlations were of expected magnitude and direction.

Support for the construct validity of QWB-SA in musculoskeletal conditions was also provided by a report of significant correlations with the scores on arthritis-specific measures (Rapid Assessment of Disease Activity in Rheumatology [RADAR], Arthritis Impact Measurement Scales, and the Health Assessment Questionnaire [HAQ]) (212). The correlations were in the low ($r = -0.28$ for QWB-SA with RADAR) to moderate ($r = -0.62$ for QWB-SA with HAQ) range. However, while correlations were in the hypothesized direction, the authors did not provide specific predictions about the strength of the expected correlations, which makes it difficult to draw robust conclusions about the convergent and discriminant validity of the QWB-SA in musculoskeletal diseases.

The QWB-SA also appears to have good ability to differentiate patients with and without musculoskeletal conditions and between severity levels of musculoskeletal conditions, further supporting its construct validity. Patients with arthritis ($n = 334$) were reported to have significantly lower QWB-SA scores and significantly higher HAQ scores than those without arthritis ($n = 562$) (212). In another study, QWB scores were sensitive to different levels of osteoarthritis severity (206), although no effect sizes (ES) for the magnitude of the differences in QWB scores for different levels of osteoarthritis severity have been provided.

Ability to detect change. Information on the ability of QWB-SA to detect change in rheumatic conditions is limited. QWB, on the other hand, was reported to be sensitive to changes in HRQOL of people with osteoarthritis following education and self-management intervention (standardized response mean 0.24) (206). In another study, the QWB also had modest ability to detect change in the health status of 330 patients with rheumatoid arthritis, with only a small standardized ES recorded (0.23) following pharmaceutical treatment, although this was similar to the ES found for other measures used in the same study, including the HAQ (ES 0.25) and tender joint count (ES 0.24) (208). There appears to be no published data on minimal clinical important differences for either the QWB or QWB-SA in either rheumatic populations or broader literature.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The QWB-SA offers comprehensive coverage of health state levels, with no compelling evidence of floor or ceiling effects, which makes this scale potentially useful across a broad range of HRQOL impairment levels. There is also good evidence for the construct validity of the QWB-SA in musculoskeletal conditions, although its appropriateness in specific disease groups and in various treatment interventions awaits further evaluations.

Caveats and cautions. Limited information is currently available on psychometric properties of the QWB-SA or QWB in musculoskeletal conditions. Since no patient samples were involved in the development or weighting of scale items, the relevance of different health states to different clinical populations is not known. In addition, the score range on QWB-SA does not allow for health states worse than death, which might make this instrument insensitive for measuring very poor health. The scoring algorithm assumes that missing responses are equivalent to absence of a problem; however, validity of this assumption is not certain. Normative scale values are only available for the US, and further studies are required to develop cross-cultural norms as well as norms for clinical populations. While there appears to be a substantial amount of evidence that support the construct validity of the QWB-SA, most psychometric evaluation studies were carried out by the scale developers, therefore further inquiries into psychometric properties of the QWB-SA by independent groups are warranted.

Clinical/research usability. Overall, the QWB-SA appears to have good support for its construct validity in rheumatic conditions, which supports its use in clinical and research settings. However, given the limited evidence for the reliability of this scale, information on its validity needs to be interpreted with caution. While the scale could potentially be useful for comparing HRQOL in rheumatic conditions with other populations (clinical or general), the complicated wording and sentence structure may limit the utility of this scale in clinical settings, where individuals may be expected to be unwell or with those who are elderly or have low levels of education. A complicated scoring system may further limit the use of the scale in clinical settings. Absence of appropriate norms and lack of information on reliability, ability to detect change, and minimal detectible change could also potentially limit the use of scale in clinical and research settings due to the difficulty with interpreting change in scale scores.

ASSEMSSMENT OF QUALITY OF LIFE (AQoL)

Description

Purpose. The AQoL instruments are multi-attribute utility measures of health-related quality of life (HRQOL) (213). In a similar way to the other utility measures (EuroQol 5-domain [EQ-5D], Short Form 6D [SF-6D], and Health Utilities Index Mark 3 [HUI3]) the AQoL was designed for use across health conditions to enable health economic evaluation studies. The AQoL allows assessment of the impact of interventions on HRQOL, comparing HRQOL in different populations and disease settings, and monitoring longitudinal changes in a broad range of health conditions. The AQoL was originally published in 1999 (214), with 4 versions developed to date: AQoL-4D (the original version), AQoL-6D (with additional elements of pain and coping), AQoL-7D (with emphasis on vision), and AQoL-8D (with emphasis on mental health) (213). This review focuses on the original version, AQoL-4D since it is the one that has been previously used in rheumatic diseases. Where possible, we also review information on the AQoL-6D due its potential relevance in rheumatology settings. A version with 8 items has also been published (215) although this version has not undergone specific validation studies in musculoskeletal conditions.

Content. The AQoL-4D covers 4 domains of independent living, mental health, relationships, and senses. The AQoL-6D has 2 additional domains of coping and pain.

Number of items. The AQoL-4D has 12 items, with 3 items per dimension. The AQoL-6D has 20 items; the additional dimensions of coping and pain have 4 items each.

Response options/scale. The AQoL items have variable numbers of response levels, ranging from 4–7. Response options are on a Guttman scale, with higher scores indicative of progressively higher levels of disability. A visual analog scale version of the AQoL is also available.

Recall period for items. The AQoL asks respondents to evaluate their health state over the previous week.

Examples of use. In rheumatology, the AQoL has been previously used in a probability sample of the general population to compare the HRQOL of people with arthritis to those who have no arthritis (216), to assess the HRQOL of people on a waiting list for joint replacement surgery (217), to evaluate the impact of self-management (218,219) and exercise-based interventions on HRQOL in arthritis (220–222), as well as in a randomized controlled trial of vertebroplasty for osteoporotic vertebral fractures (223).

Practical Application

How to obtain. The AQoL questionnaires and scoring algorithms are available at no cost from http://www.aqol.com.au/. However, the use of the AQoL is subject to copyright restrictions and the users are asked to complete a registration form (using web-based or paper format).

Method of administration. The AQoL can be self- (paper and pencil or online) or interviewer-administered. The agreement between self- and interviewer-administered (by telephone) versions of the AQoL was high with an intraclass correlation coefficient of 0.83 (95% confidence interval [95% CI] 0.76–0.88), with the 2 versions producing comparable mean scores (224). However, in another study, the correlation between mail and telephone administration of the AQoL was only 0.66, indicating that different methods of AQoL administration should not be used interchangeably (225).

Scoring. The AQoL instruments can be used to obtain an overall utility score as well as to separate scores for each dimension. The health states described between the items are initially weighted using values obtained from the general population from Time Trade Off interviews, a common procedure in the health economics field. The scores across the scales are combined using a multiplicative scoring procedure. Scoring algorithms are available from the AQoL web site (www.aqol.com) in SPSS and STATA readable formats. The AQoL developers also provide an online scoring service for their questionnaires. The scoring algorithm allows for only 1 missing value per dimension for dimensions with 3 or 4 items and 2 missing values per dimension for longer scales. Missing values are imputed from the mean of the nonmissing items in the dimension (213).

Score interpretation. The AQoL utility score ranges from −0.04 (health state worse than death) to 0.00 (death) and 1.00 (full health) (226). Normative values, broken down by age (in 10-year age groups) and sex, are available for AQoL-4D from the AQoL web site (http://www.aqol.com.au/documents/AQoL-4D-Population-Norm.pdf). The norms have been derived from a probability sample of 3,010 Australian residents (213).

Respondent burden. The AQoL has a low respondent burden. The scale developers estimate completion time for the AQoL-4D to be 1 to 2 minutes, although a more realistic estimate for a 12-item questionnaire that uses the Guttmann response scale might be 5–10 minutes, which is still quite low (213). Completion instructions are self-explanatory and easy to follow. The questionnaire uses simple language and is easy to understand and complete. The developers reported that in interview settings, ~2% of respondents tend to seek clarification about an item or a response option. Detailed information about items for which clarification is commonly sought can be found in the user manual (225), which can be downloaded from http://www.psychiatry.unimelb.edu.au/centres-units/cpro/aqol/instruments/AQoL_User_Manual.pdf. Some items describing poor HRQOL were also found to be distressing for some participants (214).

Administrative burden. The AQoL appears to have low administrator burden. Administering AQoL by interview requires basic training in interviewing technique. The use of the computerized scoring algorithm requires basic knowledge of statistical software.

Translations/adaptations. No translations or adaptations of the AQoL were identified at the time of preparing this review.

Psychometric Information

Method of development. The conceptual model for the initial version of the AQoL was based on the World Health Organization's definition of health. The 2 major sources of items for the AQoL were focus groups of clinicians and the review of the content of existing HRQOL questionnaires. No patients took part in item generation. The 61 draft items of the AQoL were administered to a sample of 255 individuals recruited from community and hospital settings. The final selection of items to be included in the AQoL-4D was made based on exploratory and confirmatory factor analysis and reliability analyses (214). The additional items for the later version of the AQoL were developed from focus groups with clinicians and review of existing questionnaires (213).

Acceptability. The AQoL appears to have high acceptability overall. In community-based studies, the proportion of missing data varies from <1% (for either self- or interviewer-administered) (224) to 2.5% (self-administered version) (226). The ability of the AQoL to adequately cover the full range of HRQOL states appears to be good in rheumatology, with no floor or ceiling effects recorded in a sample of 222 osteoarthritis patients recruited from clinical and community settings (227).

Reliability. Internal consistency (Cronbach's alpha) of the AQoL utility score is good and is generally reported to be ~0.80 in samples consisting of hospital patients and community-dwelling adults (225,226). Although the 3-item domains of the original version of the AQoL were reported to have much lower internal consistency estimates, with coefficients ranging from 0.52 (psychological well-being) to 0.77 (independent living) (214), the AQoL was intended to be used primarily as an overall utility score, rather than as single domain scores.

Information on test–retest reliability of the AQoL is currently limited. The user manual reports test–retest reliability, measured by Pearson's correlation coefficient, as 0.80 (225). However, Pearson's correlation coefficient tends to be a poor indicator of temporal stability, due to the insensitivity to systematic (rather than random) changes over time. Systematic differences in questionnaire scores over time could occur for a variety of reasons, including real change in a health state, change in internal frame of reference for the severity of one's health condition (response shift), reactivity, or learning effect. Evidence indicates that the AQoL may be subject to such systematic biases. Over repeated administrations, the AQoL-6D scores were somewhat higher for the second administration, which suggests that the individuals tend to re-apprise the severity of their condition after some reflection (228).

Validity. The AQoL instruments cover a broad range of health domains, not all of which (e.g., vision) are relevant to rheumatology. Nonetheless, these domains represent important elements of overall generic HRQOL and permit comparisons across diseases and populations. The AQoL appears to have good face and content validity for measuring HRQOL, although content validity is largely dependent upon the nature of the construct being measured. The absence of floor or ceiling effects in osteoarthritis further supports content validity of the AQoL in rheumatology since it indicates that the AQoL is able to adequately capture the full range of HRQOL experiences in this population (227). Criterion-related validity of HRQOL measures is difficult to establish due to the absence of a "gold standard" for measuring HRQOL.

Evidence for construct validity of the AQoL is good, with results thus far supporting its factorial, convergent discriminant, and known-groups validity. Factorial structure of the AQoL-4D, including the 4 first-order factors and 1 higher-order factor, was examined in the initial construction study (214) using confirmatory factor analysis, with no evidence of misfit between the hypothesized model and the data. At least one other study each subsequently supported the 4-dimensional structure of the AQoL-4D using exploratory factor analysis (229) and the 6-dimensional structure of the AQoL-6D (228). While these results provide strong support for the factorial validity of the AQoL, it should be noted that none of these studies were specifically concerned with rheumatology populations.

In rheumatology settings, convergent validity of AQoL-4D was tested in a study of 222 individuals with osteoarthritis (227), where AQoL utility had high to moderate correlations with the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) scales ($r = -0.51, -0.63$) and the Lequesne Index ($r = -0.76$). All correlations were of hypothesized magnitude and direction. More broadly, in a sample of 606 individuals drawn from community and hospital settings, correlations between the AQoL-6D and other generic measures of HRQOL, including the HUI3, EQ-5D, 15D, and the SF-36 were 0.73 or higher (228), indicating good convergent validity. The AQoL-4D utility scores also correlated well with health care costs in an 18-month followup of more than 1,500 individuals with a range of chronic conditions. While these results support convergent validity of the AQoL-4D, less is known about its discriminant validity, which needs further study.

The AQoL has good ability to differentiate between people with and without rheumatic conditions, as well as between severity levels in rheumatic conditions. In a large probability sample of the general population ($n = 2,840$), the AQoL-4D was able to differentiate people with chronic joint conditions (self-reported doctor-diagnosed arthritis and chronic joint symptoms) from those who had no joint problems, with the lowest mean AQoL scores for arthritis group (mean 0.72; 99% CI 0.70–0.74), followed by chronic joint symptoms group (mean 0.75; 99% CI 0.72–0.78), and those who had no joint problems (mean 0.85; 99% CI 0.84–0.87) (215). The AQoL-4D was also able to differentiate between severity levels of osteoarthritis, with the utility score exhibiting moderate effect size (ES) of 0.66 for the difference in HRQOL between people with osteoarthritis recruited from the general community and those who were on a waiting list for joint replacement surgery for their osteoarthritis (227). Similar results were reported in at least one other study (217). More broadly, in a sample of

996 individuals selected to cover a very broad range of health conditions from those who were healthy to those who were terminally ill, the AQoL was reported to have better ability to differentiate between the levels of HRQOL impairments than other utility instruments, including HUI3, EQ-5D, 15D, and SF6D (230).

Ability to detect change. The ability of the AQoL to detect change in rheumatic populations has not been well studied. More generally, a minimum clinically important difference (MCID) of 0.06 for the AQoL-4D utility score had been recorded for self-reported change in health state (226). This finding was based on the results of 4 longitudinal studies (with approximate followup time of 12 months), 2 of which were community trials of coordinated care for people at risk of hospitalization, 1 involved a followup of community-dwelling elderly people, and 1 was an evaluation of health services for acute conditions in a hospital emergency department. As this study did not specifically target individuals with rheumatic conditions, transferability of this finding to rheumatology settings is currently not known.

The results for the ability of the AQoL to detect treatment effects in rheumatology settings are mixed. In a randomized controlled trial of the efficacy of physiotherapy and exercise program for chronic rotator cuff disease, the mean change in AQoL-4D utility score following 22 weeks of treatment was 0.00 (SD 0.20) (220). At the same time, condition-specific measures of pain and movement (assessed using the Shoulder Pain and Disability Index) showed large improvements during the course of intervention (standardized response mean 0.90 for movement and 1.05 for pain). Nonetheless, in the same study the AQoL-4D was able to distinguish between intervention and placebo groups at 22 weeks of followup, with significantly higher scores recorded for the intervention group (mean difference 0.07; 95% CI 0.04–0.10). In a randomized controlled trial of self-management intervention for people on a waiting list for joint replacement surgery, the intervention group had a slightly higher AQoL utility score at the end of the study (ES 0.21) (218). Although the improvement was small and not statistically significant ($P = 0.23$), similar results were obtained on the WOMAC (ES 0.09, 0.36, and 0.26 for pain, stiffness, and physical functioning scales, respectively).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. As with all the HRQOL scales, the AQoL covers a range of issues important to rheumatology. The AQoL appears to have good ability to differentiate between people with and without arthritis and between the levels of arthritis severity. Overall, the evidence supports the use of the AQoL when comparisons with the general population are required. The ability of the AQoL to detect treatment effects is promising but requires further research in a broader range of interventions with treatment effects of known magnitude.

Caveats and cautions. Only a handful of studies examined the psychometric properties of the AQoL in rheumatic conditions, with generally positive results. However, the more definitive conclusions about the psychometric robustness of this questionnaire in rheumatology await further investigations.

Clinical/research usability. The AQoL is a relatively new instrument for rheumatology, and information about its psychometric properties is still accumulating. The questionnaires have low respondent and administrator burden and are available at no cost, which greatly enhances their usability in clinical and research settings. Availability of population norms also provides context for score interpretation, which further facilitates the usefulness of the AQoL. However, only Australian norms are currently available and cross-cultural applicability of these norms is currently not known. Usability of the AQoL in different countries is also affected by the lack of AQoL in languages other than English. Like all generic HRQOL tools designed to generate utilities, it is unlikely to detect small clinical changes but should be useful for comparison with other diseases and for health economic appraisals such as cost utility assessments.

DISCUSSION

The results of this review indicate that there is currently no single "best" measure of general health and health-related quality of life in rheumatology, with psychometric weaknesses identified in all measures considered. Although this review also identified several gaps in the information available on measurement properties of the reviewed questionnaires, the available evidence identifies the Sickness Impact Profile (136) as the worst performing measure, with relatively high administrative burden and questionable reliability of subscale scores. At the other end of the spectrum is the Assessment of Quality of Life Scale, with very low administrative burden and good evidence of reliability and validity thus far, indicating that it is a promising measure. The results of this review suggest that there is an urgent need for systematic investigations of the psychometric properties of many instruments currently used to assess health and health-related quality of life in rheumatology.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36) I. Conceptual framework and item selection. Med Care 1992; 30:473–83.
2. Ware J Jr, Kosinski M, Keller SD. A 12-item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220–33.
3. Ware JE. SF-36 Health Survey update. In: Maruish ME, editor. The use of psychological testing for treatment planning and outcomes assessment. Mahwah (NJ): Lawrence Earlbaum; 2004. p. 693–718.
4. Ware JE, Kosinski MA, Gandek B. SF-36 Health Survey: manual and interpretation guide. Lincoln (RI): Quality Metric; 2005.
5. Ware JE Jr, Gandek B, Kosinski M, Aaronson NK, Apolone G, Brazier J, et al. The equivalence of SF-36 summary health scores estimated using standard and country-specific algorithms in 10 countries: results

from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1167–70.

6. Hawthorne G, Osborne RH, Taylor A, Sansoni J. The SF-36 version 2: critical analyses of population weights, scoring algorithms and population norms. Qual Life Res 2007;16:661–73.
7. Boardman DL, Dorey F, Thomas BJ, Lieberman JR. The accuracy of assessing total hip arthroplasty outcomes: a prospective correlation study of walking ability and 2 validated measurement devices. J Arthroplasty 2000;15:200–4.
8. Dass S, Bowman SJ, Vital EM, Ikeda K, Pease CT, Hamburger J, et al. Reduction of fatigue in Sjögren's syndrome with rituximab: results of a randomised, double-blind, placebo-controlled pilot study. Ann Rheum Dis 2008;67:1541–4.
9. Gladman DD, Urowitz MB, Gough J, MacKinnon A. Fibromyalgia is a major contributor to quality of life in lupus. J Rheumatol 1997;24: 2145–8.
10. Harrison MJ, Tricker KJ, Davies L, Hassell A, Dawes P, Scott DL, et al. The relationship between social deprivation, disease outcome measures, and response to treatment in patients with stable, long-standing rheumatoid arthritis. J Rheumatol 2005;32:2330–6.
11. Soderlin MK, Lindroth Y, Turesson C, Jacobsson LT. A more active treatment has profound effects on the health status of rheumatoid arthritis (RA) patients: results from a population-based RA register in Malmo, Sweden, 1997-2005. Scand J Rheumatol 2010;39:206–11.
12. Heiberg MS, Nordvag BY, Mikkelsen K, Rodevand E, Kaufmann C, Mowinckel P, et al. The comparative effectiveness of tumor necrosis factor-blocking agents in patients with rheumatoid arthritis and patients with ankylosing spondylitis: a six-month, longitudinal, observational, multicenter study. Arthritis Rheum 2005;52:2506–12.
13. Becker MA, Schumacher HR, Benjamin KL, Gorevic P, Greenwald M, Fessel J, et al. Quality of life and disability in patients with treatment-failure gout. J Rheumatol 2009;36:1041–8.
14. Busija L, Osborne RH, Nilsdotter A, Buchbinder R, Roos EM. Magnitude and meaningfulness of change in SF-36 scores in four types of orthopedic surgery. Health Qual Life Outcomes 2008;6:55.
15. Dervin GF, Stiell IG, Rody K, Grabowski J. Effect of arthroscopic debridement for osteoarthritis of the knee on health-related quality of life. J Bone Joint Surg Am 2003;85A:10–9.
16. Dierick F, Aveniere T, Cossement M, Poilvache P, Lobet S, Detrembleur C. Outcome assessment in osteoarthritic patients undergoing total knee arthroplasty. Acta Orthop Belg 2004;70:38–45.
17. Maini RN, Breedveld FC, Kalden JR, Smolen JS, Furst D, Weisman MH, et al, Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. Sustained improvement over two years in physical function, structural damage, and signs and symptoms among patients with rheumatoid arthritis treated with infliximab and methotrexate. Arthritis Rheum 2004;50:1051–65.
18. Linde K, Weidenhammer W, Streng A, Hoppe A, Melchart D. Acupuncture for osteoarthritic pain: an observational study in routine care. Rheumatology (Oxford) 2006;45:222–7.
19. Ferrara PE, Rabini A, Maggi L, Piazzini DB, Logroscino G, Magliocchetti G, et al. Effect of pre-operative physiotherapy in patients with end-stage osteoarthritis undergoing hip arthroplasty. Clin Rehabil 2008;22:977–86.
20. Sutbeyaz ST, Sezer N, Koseoglu F, Kibar S. Low-frequency pulsed electromagnetic field therapy in fibromyalgia: a randomized, double-blind, sham-controlled clinical study. Clin J Pain 2009;25:722–8.
21. Wang C, Schmid CH, Hibberd PL, Kalish R, Roubenoff R, Rones R, et al. Tai Chi is effective in treating knee osteoarthritis: a randomized controlled trial. Arthritis Rheum 2009;61:1545–53.
22. Coleman S, Briffa NK, Carroll G, Inderjeeth C, Cook N, McQuade J. Effects of self-management, education and specific exercises, delivered by health professionals, in patients with osteoarthritis of the knee. BMC Musculoskelet Disord 2008;9:133.
23. Carmona L, Ballina J, Gabriel R, Laffon A. The burden of musculoskeletal diseases in the general population of Spain: results from a national survey. Ann Rheum Dis 2001;60:1040–5.
24. Hill CL, Gill T, Taylor AW, Daly A, Grande ED, Adams RJ. Psychological factors and quality of life in arthritis: a population-based study. Clin Rheumatol 2007;26:1049–54.
25. Gandhi SK, Salmon JW, Zhao SZ, Lambert BL, Gore PR, Conrad K. Psychometric evaluation of the 12-item Short-Form Health Survey (SF-12) in osteoarthritis and rheumatoid arthritis clinical trials. Clin Ther 2001;23:1080–98.
26. Theiler R, Bischoff HA, Good M, Uebelhart D. Rofecoxib improves quality of life in patients with hip or knee osteoarthritis. Swiss Med Wkly 2002;132:566–73.
27. Foley A, Halbert J, Hewitt T, Crotty M. Does hydrotherapy improve strength and physical function in patients with osteoarthritis: a randomised controlled trial comparing a gym based and a hydrotherapy based strengthening programme. Ann Rheum Dis 2003;62:1162–7.
28. Calandre EP, Rodriguez-Claro ML, Rico-Villademoros F, Vilchez JS, Hidalgo J, Delgado-Rodriguez A. Effects of pool-based exercise in fibromyalgia symptomatology and sleep quality: a prospective randomised comparison between stretching and Ai Chi. Clin Exp Rheumatol 2009;27:S21–8.
29. Fransen M, Nairn L, Winstanley J, Lam P, Edmonds J. Physical activity for osteoarthritis management: a randomized controlled clinical trial evaluating hydrotherapy or Tai Chi classes. Arthritis Rheum 2007;57: 407–14.
30. McCalden RW, MacDonald SJ, Rorabeck CH, Bourne RB, Chess DG, Charron KD. Wear rate of highly cross-linked polyethylene in total hip arthroplasty: a randomized controlled trial. J Bone Joint Surg Am 2009;91:773–82.
31. Thomas S, Kinninmonth AW, Kumar CS. Long-term results of the modified Hoffman procedure in the rheumatoid forefoot: surgical technique. J Bone Joint Surg Am 2006;88 Suppl 1:149–57.
32. Schick M, Stucki G, Rodriguez M, Meili EO, Huber E, Michel BA, et al. Haemophilic arthropathy: assessment of quality of life after total knee arthroplasty. Clin Rheumatol 1999;18:468–72.
33. Stockl KM, Shin JS, Lew HC, Zakharyan A, Harada AS, Solow BK, et al. Outcomes of a rheumatoid arthritis disease therapy management program focusing on medication adherence. J Manag Care Pharm 2010;16:593–604.
34. Bowling A, Bond M, Jenkinson C, Lamping DL. Short Form 36 (SF-36) Health Survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, the Health Survey for England and the Oxford Healthy Life Survey. J Public Health Med 1999;21:255–70.
35. McHorney CA, Kosinski M, Ware JE Jr. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. Med Care 1994; 32:551–67.
36. Lyons RA, Wareham K, Lucas M, Price D, Williams J, Hutchings HA. SF-36 scores vary by method of administration: implications for study design. J Public Health Med 1999;21:41–5.
37. Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? J Clin Epidemiol 1996;49:135–40.
38. Jones D, Kazis L, Lee A, Rogers W, Skinner K, Cassar L, et al. Health status assessments using the Veterans SF-12 and SF-36: methods for evaluating otucomes in the Veterans Health Administration. J Ambul Care Manage 2001;24:68–86.
39. Perkins JJ, Sanson-Fisher RW. An examination of self- and telephone-administered modes of administration for the Australian SF-36. J Clin Epidemiol 1998;51:969–73.
40. Millard RW, Carver JR. Cross-sectional comparison of live and interactive voice recognition administration of the SF-12 health status survey. Am J Manag Care 1999;5:153–9.
41. Ball AE, Russell EM, Seymour DG, Primrose WR, Garratt AM. Problems in using health survey questionnaires in older patients with physical disabilities: can proxies be used to complete the SF-36? Gerontology 2001;47:334–40.
42. Yip JY, Wilber KH, Myrtle RC, Grazman DN. Comparison of older adult subject and proxy responses on the SF-36 health-related quality of life instrument. Aging Ment Health 2001;5:136–42.
43. Kosinski M, Bayliss M, Bjorner JB, Ware JE. Improving estimates of SF-36-1 Health Survey scores for respondents with missing data. Med Outcome Trust Mon 2000;5:8–10.
44. Ware JE Jr, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Med Care 1995;33:AS264–79.
45. Jenkinson C, Coulter A, Wright L. Short form 36 (SF-36) health survey questionnaire: normative data for adults of working age. BMJ 1993; 306:1437–40.
46. Watson EK, Firman DW, Baade PD, Ring I. Telephone administration of the SF-36 health survey: validation studies and population norms for adults in Queensland. Aust N Z J Public Health 1996;20:359–63.
47. Australian Bureau of Statistics. National health survey: SF36 population norms, Australia, 1995. Cat. no. 4399.0. Canberra (Australia): ABS; 1997.
48. Sullivan M, Karlsson J, Ware JE. SF-36 Swedish manual and interpretation guide. Gothenburg: Gothenburg University; 1994.
49. Thumboo J, Chan SP, Machin D, Soh CH, Feng PH, Boey ML, et al. Measuring health-related quality of life in Singapore: normal values for the English and Chinese SF-36 health survey. Ann Acad Med Singapore 2002;31:366–74.
50. Scott KM, Tobias MI, Sarfati D, Haslett SJ. SF-36 health survey reliability, validity and norms for New Zealand. Aust N Z J Public Health 1999;23:401–6.
51. Hanmer J, Lawrence WF, Anderson JP, Kaplan RM, Fryback DG. Report of nationally representative values for the noninstitutionalized US adult population for 7 health-related quality-of-life scores. Med Decis Making 2006;26:391–400.

52. Ware JE, Kosinski MA, Turner-Bowker DM, Gandek B. SF-12: how to score version 2 of the SF-12 Health Survey (with a supplement documenting version 1). Lincoln (RI): QualityMetric; 2002.
53. Gandek B, Ware JE, Aaronson NK, Apolone G, Bjorner JB, Brazier JE, et al. Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51: 1171–8.
54. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000;17:13–35.
55. Parker SG, Bechinger-English D, Jagger C, Spiers N, Lindesay J. Factors affecting completion of the SF-36 in older people. Age Ageing 2006; 35:376–81.
56. Adamson J, Gooberman-Hill R, Woolhead G, Donovan J. 'Questerviews': using questionnaires in qualitative interviews as a method of integrating qualitative and quantitative health services research. J Health Sci Res Pol 2004;9:139–45.
57. Fowler RW, Congdon P, Hamilton S. Assessing health status and outcomes in a geriatric day hospital. Public Health 2000;114:440–5.
58. DeBrota DJ, Bradt EW, Andrejasich CM, Kosinski M, Ware JE. Comparison of interactive voice response SF-36 to self-administered SF-36 and personal interview via telephone SF-36: Boston: The Health Institute, New England Medical Center; 1996.
59. Sanson-Fisher RW, Perkins JJ. Adaptation and validation of the SF-36 Health Survey for use in Australia. J Clin Epidemiol 1998;51:961–7.
60. Tarlov AR, Ware JE Jr, Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study: an application of methods for monitoring the results of medical care. JAMA 1989;262:925–30.
61. Koh ET, Leong KP, Tsou IY, Lim VH, Pong LY, Chong SY, et al. The reliability, validity and sensitivity to change of the Chinese version of SF-36 in oriental patients with rheumatoid arthritis. Rheumatology (Oxford) 2006;45:1023–8.
62. Loge JH, Kaasa S, Hjermstad MJ, Kvien TK. Translation and performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. I. Data quality, scaling assumptions, reliability, and construct validity. J Clin Epidemiol 1998;51:1069–76.
63. Parker SG, Peet SM, Jagger C, Farhan M, Castleden CM. Measuring health status in older patients: the SF-36 in practice. Age Ageing 1998;27:13–8.
64. Linde L, Sorensen J, Ostergaard M, Horslev-Petersen K, Rasmussen C, Jensen DV, et al. What factors influence the health status of patients with rheumatoid arthritis measured by the SF-12v2 Health Survey and the Health Assessment Questionnaire? J Rheumatol 2009;36:2183–9.
65. Soderman P, Malchau H. Validity and reliability of Swedish WOMAC osteoarthritis index: a self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP). Acta Orthop Scand 2000;71:39–46.
66. Saleh KJ, Radosevich DM, Kassim RA, Moussa M, Dykes D, Bottolfson H, et al. Comparison of commonly used orthopaedic outcome measures using palm-top computers and paper surveys. J Orthop Res 2002;20:1146–51.
67. Kvien TK, Kaasa S, Smedstad LM. Performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. II. A comparison of the SF-36 with disease-specific measures. J Clin Epidemiol 1998;51:1077–86.
68. Linde L, Sorensen J, Ostergaard M, Horslev-Petersen K, Hetland ML. Health-related quality of life: validity, reliability, and responsiveness of SF-36, 15D, EQ-5D (corrected) RAQoL, and HAQ in patients with rheumatoid arthritis. J Rheumatol 2008;35:1528–37.
69. Kosinski M, Keller SD, Hatoum HT, Kong SX, Ware JE Jr. The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: tests of data quality, scaling assumptions and score reliability. Med Care 1999;37: MS10–22.
70. Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36). Br J Rheumatol 1998;37:425–36.
71. Davey RC, Edwards SM, Cochrane T. Test-retest reliability of lower extremity functional and self-reported measures in elderly with osteoarthritis. Adv Physiother 2003;5:155–61.
72. Brazier JE, Harper R, Munro J, Walters SJ, Snaith ML. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. Rheumatology (Oxford) 1999;38:870–7.
73. Jakobsson U, Westergren A, Lindskov S, Hagell P. Construct validity of the SF-12 in three different samples. J Eval Clin Pract. E-pub ahead of print.
74. Lenert LA. The reliability and internal consistency of an Internet-capable computer program for measuring utilities. Qual Life Res 2000; 9:811–7.
75. Lim LL, Fisher JD. Use of the 12-item Short-Form (SF-12) Health Survey in an Australian heart and stroke population. Qual Life Res 1999;8:1–8.
76. Luo X, Lynn George M, Kakouras I, Edwards CL, Pietrobon R, Richardson W, et al. Reliability, validity, and responsiveness of the short form 12-item survey (SF-12) in patients with back pain. Spine (Phila Pa 1976) 2003;28:1739–45.
77. Resnick B, Nahm ES. Reliability and validity testing of the revised 12-item Short-Form Health Survey in older adults. J Nurs Meas 2001; 9:151–61.
78. Salyers MP, Bosworth HB, Swanson JW, Lamb-Pagone J, Osher FC. Reliability and validity of the SF-12 health survey among people with severe mental illness. Med Care 2000;38:1141–50.
79. Keller SD, Ware JE Jr, Bentler PM, Aaronson NK, Alonso J, Apolone G, et al. Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1179–88.
80. Wolinsky FD, Stump TE. A measurement model of the Medical Outcomes Study 36-Item Short-Form Health Survey in a clinical sample of disadvantaged, older, black, and white men and women. Med Care 1996;34:537–48.
81. Shields RK, Enloe LJ, Leo KC. Health related quality of life in patients with total hip or knee replacement. Arch Phys Med Rehabil 1999;80: 572–9.
82. Kontodimopoulos N, Pappa E, Niakas D, Tountas Y. Validity of SF-12 summary scores in a Greek general population. Health Qual Life Outcomes 2007;5:55.
83. Montazeri A, Vahdaninia M, Mousavi SJ, Omidvari S. The Iranian version of 12-item Short Form Health Survey (SF-12): factor structure, internal consistency and construct validity. BMC Public Health 2009; 9:341.
84. Maurischat C, Ehlebracht-Konig I, Kuhn A, Bullinger M. Factorial validity and norm data comparison of the Short Form 12 in patients with inflammatory-rheumatic disease. Rheumatol Int 2006;26:614–21.
85. Fleishman JA, Selim AJ, Kazis LE. Deriving SF-12v2 physical and mental health summary scores: a comparison of different scoring algorithms. Qual Life Res 2010;19:231–41.
86. Veehof MM, ten Klooster PM, Taal E, van Riel PL, van de Laar MA. Comparison of internal and external responsiveness of the generic Medical Outcome Study Short Form-36 (SF-36) with disease-specific measures in rheumatoid arthritis. J Rheumatol 2008;35:610–7.
87. Fan ZJ, Smith CK, Silverstein BA. Assessing validity of the QuickDASH and SF-12 as surveillance tools among workers with neck or upper extremity musculoskeletal disorders. J Hand Ther 2008;21:354–65.
88. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. J Epidemiol Comm Health 1980;34:281–6.
89. Jenkinson C, Fitzpatrick R, Argyle M. The Nottingham Health Profile: an analysis of its sensitivity in differentiating illness groups. Soc Sci Med 1988;27:1411–4.
90. Hunt SM, McKenna SP, McEwen J, Williams J, Papp E. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med A 1981;15:221–9.
91. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract 1985;35: 185–8.
92. Baillet A, Payraud E, Niderprim VA, Nissen MJ, Allenet B, François P, et al. A dynamic exercise programme to improve patients' disability in rheumatoid arthritis: a prospective randomized controlled trial. Rheumatology (Oxford) 2009;48:410–5.
93. Ekici G, Bakar Y, Akbayrak T, Yuksel I. Comparison of manual lymph drainage therapy and connective tissue massage in women with fibromyalgia: a randomized controlled trial. J Manipulative Physiol Ther 2009;32:127–33.
94. Yurtkuran M, Alp A, Nasircilar A, Bingol U, Altan L, Sarpdere G. Balneotherapy and tap water therapy in the treatment of knee osteoarthritis. Rheumatol Int 2006;27:19–27.
95. Evcik D, Kavuncu V, Yeter A, Yigit I. The efficacy of balneotherapy and mud-pack therapy in patients with knee osteoarthritis. Joint Bone Spine 2007;74:60–5.
96. Knahr K, Korn V, Kryspin-Exner I, Jagsch R. Quality of life five years after total or partial knee arthroplasty. Z Orthop Ihre Grenzgeb 2003; 141:27–32. In German.
97. Atroshi I, Ornstein E, Franzen H, Johnsson R, Stefansdottir A, Sundberg M. Quality of life after hip revision with impaction bone grafting on a par with that 4 years after primary cemented arthroplasty. Acta Orthop Scand 2004;75:677–83.
98. Uutela T, Hannonen P, Kautiainen H, Hakala M, Paananen ML, Hakkinen A. Positive treatment response improves the health-related quality of life of patients with early rheumatoid arthritis. Clin Exp Rheumatol 2009;27:108–11.

99. Wigers SH, Finset A. Rehabilitation of chronic myofascial pain disorders. Tidsskr Nor Laegeforen 2007;127:604–8. In Norwegian.
100. Kersten P, Mullee MA, Smith JA, McLellan L, George S. Generic health status measures are unsuitable for measuring health status in severely disabled people. Clin Rehabil 1999;13:219–28.
101. Vidalis A, Syngelakis M, Papathanasiou M, Whalley D, McKenna SP. The Greek version of the Nottingham Health Profile: features of its adaptation. Hippokratia 2002;6 Suppl 1:75–8.
102. Bucquet D, Condon S, Ritchie K. The French version of the Nottingham Health Profile: a comparison of items weights with those of the source version. Soc Sci Med 1990;30:829–35.
103. Wiklund I, Romanus B, Hunt SM. Self-assessed disability in patients with arthrosis of the hip joint: reliability of the Swedish version of the Nottingham Health Profile. Disabil Rehabil 1988;10:159–63.
104. Essink-Bot ML, Krabbe PF, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures: the Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. Med Care 1997;35:522–37.
105. Alonso J, Prieto L, Anto JM. The Spanish version of the Nottingham Health Profile: a review of adaptation and instrument characteristics. Qual Life Res 1994;3:385–93.
106. Hunt SM, McEwen J, McKenna SP. Measuring health status. London: Croom Helm; 1986.
107. Post MW, Gerritsen J, Diederiks JP, De Witte LP. Measuring health status of people who are wheelchair-dependent: validity of the sickness impact profile 68 and the Nottingham health profile. Disabil Rehabil 2001;23:245–53.
108. Nagyova I, van den Heuvel W, Steward R, Macejova Z, van Dijk J. Predictors of change in self-rated health: a longitudinal analysis in patients with rheumatoid arthritis. Netherlands: University of Groningen; 2005.
109. VanderZee KI, Sanderman R, Heyink J. A comparison of two multidimensional measures of health status: the Nottingham Health Profile and the RAND 36-Item Health Survey 1.0. Qual Life Res 1996;5: 165–74.
110. Hunt SM, McKenna SP, Williams J. Reliability of a population survey tool for measuring perceived health problems: a study of patients with osteoarthrosis. J Epidemiol Comm Health 1981;35:297–300.
111. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79–93.
112. Bouchet C, Guillemin F, Briancon S. Comparison of 3 quality of life instruments in the longitudinal study of rheumatoid arthritis. Rev Epidemiol Sante Publique 1995;43:250–8. In French.
113. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A generic health status instrument in the assessment of rheumatoid arthritis. Br J Rheumatol 1992;31:87–90.
114. Houssien DA, McKenna SP, Scott DL. The Nottingham Health Profile as a measure of disease activity and outcome in rheumatoid arthritis. Br J Rheumatol 1997;36:69–73.
115. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. Importance of sensitivity to change as a criterion for selecting health status measures. Qual Health Care 1992;1:89–93.
116. Bachrach-Lindstrom M, Karlsson S, Pettersson LG, Johansson T. Patients on the waiting list for total hip replacement: a 1-year follow-up study. Scand J Caring Sci 2008;22:536–42.
117. Lillegraven S, Kvien TK. Measuring disability and quality of life in established rheumatoid arthritis. Best Pract Res Clin Rheumatol 2007; 21:827–40.
118. De Bruin AF, Buys M, De Witte LP, Diederiks JP. The sickness impact profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility. J Clin Epidemiol 1994;47:863–71.
119. Gilson BS, Gilson JS, Bergner M, Bobbit RA, Kressel S, Pollard WE, et al. The sickness impact profile: development of an outcome measure of health care. Am J Public Health 1975;65:1304–10.
120. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: development and final revision of a health status measure. Med Care 1981;19:787–805.
121. De Bruin AF, Diederiks JP, De Witte LP, Stevens FC, Philipsen H. The development of a short generic version of the sickness impact profile. J Clin Epidemiol 1994;47:407–18.
122. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8:141–4.
123. Sullivan M, Ahlmen M, Bjelle A, Karlsson J. Health status assessment in rheumatoid arthritis. II. Evaluation of a modified Shorter Sickness Impact Profile. J Rheumatol 1993;20:1500–7.
124. Knutsson S, Engberg IB. An evaluation of patients' quality of life before, 6 weeks and 6 months after total hip replacement surgery. J Adv Nurs 1999;30:1349–59.
125. Tak E, Staats P, Van Hespen A, Hopman-Rock M. The effects of an exercise program for older adults with osteoarthritis of the hip. J Rheumatol 2005;32:1106–13.
126. Ahlmen M, Sullivan M, Bjelle A. Team versus non-team outpatient care in rheumatoid arthritis: a comprehensive outcome evaluation including an overall health measure. Arthritis Rheum 1988;31:471–9.
127. SIP (Sickness Impact Profile). Mapi Research Trust Education Information Dissemination, 2010. URL: http://www.mapi-trust.org/services/questionnairelicensing/cataloguequestionnaires/118-sip.
128. Agel J, Swiontkowski MF. Guide to outcomes instruments for musculoskeletal trauma research. J Orthop Trauma 2006;20:S1–146.
129. De Bruin AF, De Witte LP, Stevens F, Diederiks JP. Sickness impact profile: the state of the art of a generic functional status measure. Soc Sci Med 1992;35:1003–14.
130. Gilson BS, Bergner M, Bobbitt RA, Carter WB. The Sickness Impact Profile: final development and testing. Hyattsville (MD): National Center for Health Services Research; 1979.
131. Rothman ML, Hedrick S, Inui T. The Sickness Impact Profile as a measure of the health status of noncognitively impaired nursing home residents. Med Care 1989;27:S157–67.
132. Anonymous. The Sickness Impact Profile 68 (SIP 68). Spinal Cord Injury Rehabilitation Evidence 2010. URL: http://www.scireproject.com/outcome-measures/sickness-impact-profile-68-sip-68.
133. Sickness Impact Profile (SIP). 2010. URL: http://www.proqolid.org/instruments/sickness_impact_profile_sip.
134. Patrick D, Peach H, editors. Disablement in the community. Oxford: Oxford University Press; 1989.
135. Bergner M, Bobbitt RA, Kressel S. The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. Int J Health Serv 1976;6:393–415.
136. Andresen EM, Rothenberg BM, Kaplan RM. Performance of a self-administered mailed version of the quality of well-being (QWB-SA) questionnaire among older adults. Med Care 1998;36:1349–60.
137. Van Straten A, De Haan RJ, Limburg M, Schuling J, Bossuyt PM, van den Bos GA. A stroke-adapted 30-item version of the sickness impact profile to assess quality of life (SA-SIP30). Stroke 1997;28:2155–61.
138. Nanda U, McLendon PM, Andresen EM, Armbrecht E. The SIP68: an abbreviated sickness impact profile for disability outcomes research. Qual Life Res 2003;12:583–95.
139. Kessler S, Jaeckel W, Cziske R. Assessing health in musculoskeletal disorders: the appropriateness of a German version of the Sickness Impact Profile. Rheumatol Int 1997;17:119–25.
140. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Control Clin Trials 1991;12:142S–58S.
141. Pollard WE, Bobbitt RA, Bergner M, Martin DP, Gilson BS. The Sickness Impact Profile: reliability of a health status measure. Med Care 1976;14:146–55.
142. Andresen EM, Nanda U, McLendon P, Meyer A, Armbrech E. SIP68: an abbreviated Sickness Impact Profile for disability outcomes research? [abstract]. Qual Life Res 2000;9:343.
143. Weinberger M, Samsa GP, Tierney WM, Belyea MJ, Hiner SL. Generic versus disease specific health status measures: comparing the Sickness Impact Profile and the Arthritis Impact Measurement Scales. J Rheumatol 1992;19:543–6.
144. Deyo RA, Patrick DL. The significance of treatment effects: the clinical perspective. Med Care 1995;33:AS286–91.
145. Post MW, Gerritsen J, van Leusen ND, Paping MA, Prevo AJ. Adapting the Nottingham Health Profile for use in people with severe physical disabilities. Clin Rehabil 2001;15:103–10.
146. Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. Health Serv Res 1984;19:275–89.
147. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992;30:917–25.
148. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632–42.
149. Bryan S, Longworth L. Measuring health-related utility: why the disparity between EQ-5D and SF-6D? Eur J Health Econ 2005;6:253–60.
150. Grieve R, Grishchenko M, Cairns J. SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility. Eur J Health Econ 2009;10:15–23.
151. Hawthorne G, Richardson J. Measuring the value of program outcomes: a review of multiattribute utility measures. Expert Rev Pharmacoeconomics Outcome Res 2001;1:215–28.
152. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005;14:1523–32.
153. Wolfe F, Michaud K, Wallenstein G. Scale characteristics and mapping accuracy of the US EQ-5D, UK EQ-5D, and SF-6D in patients with rheumatoid arthritis. J Rheumatol 2010;37:1615–25.
154. Joore M, Brunenberg D, Nelemans P, Wouters E, Kuijpers P, Honig A, et al. The impact of differences in EQ-5D and SF-6D utility scores on

the acceptability of cost-utility ratios: results across five trial-based cost-utility studies. Value Health 2010;13:222–9.

155. Konerding U, Moock J, Kohlmann T. The classification systems of the EQ-5D, the HUI II and the SF-6D: what do they have in common? Qual Life Res 2009;18:1249−61.

156. Marra CA, Marion SA, Guh DP, Najafzadeh M, Wolfe F, Esdaile JM, et al. Not all "quality-adjusted life years" are equal. J Clin Epidemiol 2007;60:616−24.

157. Osnes-Ringen H, Kvamme MK, Kristiansen IS, Thingstad M, Henriksen JE, Kvien TK, et al. Cost-effectiveness analyses of elective orthopaedic surgical procedures in patients with inflammatory arthropathies. Scand J Rheumatol 2011;2:108−15.

158. Barton GR, Sach TH, Avery AJ, Doherty M, Jenkinson C, Muir KR. Comparing the performance of the EQ-5D and SF-6D when measuring the benefits of alleviating knee pain [abstract]. Cost Eff Resour Alloc 2009;7:12.

159. Barton GR, Sach TH, Doherty M, Avery AJ, Jenkinson C, Muir KR. An assessment of the discriminative ability of the EQ-5Dindex, SF-6D, and EQ VAS, using sociodemographic factors and clinical conditions. Eur J Health Econ 2008;9:237−49.

160. Sach TH, Barton GR, Jenkinson C, Doherty M, Avery AJ, Muir KR. Comparing cost-utility estimates: does the choice of EQ-5D or SF-6D matter? Med Care 2009;47:889−94.

161. Ruchlin HS, Insinga RP. A review of health-utility data for osteoarthritis: implications for clinical trial-based evaluation. Pharmacoeconomics 2008;26:925–35.

162. Seymour J, McNamee P, Scott A, Tinelli M. Shedding new light onto the ceiling and floor? A quantile regression approach to compare EQ-5D and SF-6D responses. Health Econ 2010;19:683–96.

163. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271–92.

164. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care 2004;42:851–9.

165. Boonen A, Patel V, Traina S, Chiou CF, Maetzel A, Tsuji W. Rapid and sustained improvement in health-related quality of life and utility for 72 weeks in patients with ankylosing spondylitis receiving etanercept. J Rheumatol 2008;35:662–7.

166. Zijlstra TR, Braakman-Jansen LM, Taal E, Rasker JJ, van de Laar MA. Cost-effectiveness of spa treatment for fibromyalgia: general health improvement is not for free. Rheumatology (Oxford) 2007;46:1454−9.

167. Van den Hout WB, de Jong Z, Munneke M, Hazes JM, Breedveld FC, Vliet Vlieland TP. Cost-utility and cost-effectiveness analyses of a long-term, high-intensity exercise program compared with conventional physical therapy in patients with rheumatoid arthritis. Arthritis Rheum 2005;53:39−47.

168. Uhlig T, Loge JH, Kristiansen IS, Kvien TK. Quantification of reduced health-related quality of life in patients with rheumatoid arthritis compared to the general population. J Rheumatol 2007;34:1241–7.

169. Fryback DG, Dunham NC, Palta M, Hanmer J, Buechner J, Cherepanov D, et al. US norms for six generic health-related quality-of-life indexes from the National Health Measurement study. Med Care 2007;45: 1162–70.

170. Furlong W, Feeny D, Torrance GW, Barr R, Horsman J. Guide to design and development of health state utility instrumentation. Hamilton (Ontario): Centre for Health Economics and Policy Analysis, McMaster University; 1990.

171. Barton GR, Sach TH, Avery AJ, Jenkinson C, Doherty M, Whynes DK, et al. A comparison of the performance of the EQ-5D and SF-6D for individuals aged >or= 45 years. Health Econ 2008;17:815–32.

172. Slobogean GP, Noonan VK, O'Brien PJ. The reliability and validity of the Disabilities of Arm, Shoulder, and Hand, EuroQol-5D, Health Utilities Index, and Short Form-6D outcome instruments in patients with proximal humeral fractures. J Shoulder Elbow Surg 2010;19: 342−8.

173. Khanna D, Furst DE, Wong WK, Tsevat J, Clements PJ, Park GS, et al. Reliability, validity, and minimally important differences of the SF-6D in systemic sclerosis. Qual Life Res 2007;16:1083–92.

174. Boonen A, van der Heijde D, Landewe R, van Tubergen A, Mielants H, Dougados M, et al. How do the EQ-5D, SF-6D and the well-being rating scale compare in patients with ankylosing spondylitis? Ann Rheum Dis 2007;66:771–7.

175. Harrison MJ, Davies LM, Bansback NJ, McCoy MJ, Farragher TM, Verstappen SM, et al. Why do patients with inflammatory arthritis often score states "worse than death" on the EQ-5D? An investigation of the EQ-5D classification system. Value Health 2009;12:1026−34.

176. Kontodimopoulos N, Pappa E, Papadopoulos AA, Tountas Y, Niakas D. Comparing SF-6D and EQ-5D utilities across groups differing in health status. Qual Life Res 2009;18:87–97.

177. Goncalves Campolina A, Bruscato Bortoluzzo A, Bosi Ferraz M, Mesquita Ciconelli R. Validity of the SF-6D index in Brazilian patients with rheumatoid arthritis. Clin Exp Rheumatol 2009;27:237−45.

178. Aggarwal R, Wilke CT, Pickard AS, Vats V, Mikolaitis R, Fogg L, et al. Psychometric properties of the EuroQol-5D and Short Form-6D in patients with systemic lupus erythematosus. J Rheumatol 2009;36: 1209−16.

179. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60:1571−82.

180. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13: 873−84.

181. Petrou S, Hockley C. An investigation into the empirical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population. Health Econ 2005;14:1169−89.

182. Harrison MJ, Davies LM, Bansback NJ, McCoy MJ, Verstappen SM, Watson K, et al. The comparative responsiveness of the EQ-5D and SF-6D to change in patients with inflammatory arthritis. Qual Life Res 2009;18:1195–205.

183. Adams R, Walsh C, Veale D, Bresnihan B, FitzGerald O, Barry M. Understanding the relationship between the EQ-5D, SF-6D, HAQ and disease activity in inflammatory arthritis. Pharmacoeconomics 2010; 28:477−87.

184. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index(HUI®): concepts, measurement properties and applications. Health Qual Life Outcome 2003;1:54.

185. The Health Utilities Index. 2010. URL: http://www.healthutilities.com/.

186. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. Oper Res 1982;30:1043−69.

187. Boyle M, Furlong W, Torrance G, Feeny D. Reliability of the Health Utilities Index-mark III used in the 1991 cycle 6 general social survey health questionnaire. Ontario: Center for Health Economics and Policy Analysis; 1994.

188. Strand V, Singh JA. Newer biological agents in rheumatoid arthritis: impact on health-related quality of life and productivity. Drugs 2010; 70:121−45.

189. Mittendorf T, Dietz B, Sterz R, Kupper H, Cifaldi MA, von der Schulenburg JM. Improvement and longterm maintenance of quality of life during treatment with adalimumab in severe rheumatoid arthritis. J Rheumatol 2007;34:2343–50.

190. Prince FH, Geerdink LM, Borsboom GJ, Twilt M, van Rossum MA, Hoppenreijs EP, et al. Major improvements in health-related quality of life during the use of etanercept in patients with previously refractory juvenile idiopathic arthritis. Ann Rheum Dis 2010;69:138−42.

191. Raynauld JP, Torrance GW, Band PA, Goldsmith CH, Tugwell P, Walker V, et al. A prospective, randomized, pragmatic, health outcomes trial evaluating the incorporation of hylan G-F 20 into the treatment paradigm for patients with knee osteoarthritis (Part 1 of 2): clinical results. Osteoarthritis Cartilage 2002;10:506−17.

192. Cadman D, Goldsmith C, Torrance G, Boyle M, Furlong W. Development of a health status index for Ontario children. Hamilton (Ontario): McMaster University, Centre for Health Economics and Policy Analysis; 1986. Final Report to Ontario Ministry of Health Research, grant DM648: (00633).

193. Feeny DH, Torrance GW, Furlong WJ. Health Utilities Index. In: Bert Spilker, editor. Quality of life and pharmacoeconomics in clinical trials: part 2. Vol. 26. Philadelphia: Lippincott-Raven Press; 1996. pp. 239−52.

194. Luo N, Chew LH, Fong KY, Koh DR, Ng SC, Yoon KH, et al. A comparison of the EuroQol-5D and the Health Utilities Index mark 3 in patients with rheumatic disease. J Rheumatol 2003;30:2268−74.

195. Ruiz MA, Rejas J, Soto J, Pardo A, Rebollo I. Adaptation and validation of the Health Utilities Index Mark 3 into Spanish and correction norms for Spanish population. Qual Life Res 2002;11:12−8.

196. Pressler SJ, Eckert GJ, Morrison GC, Murray MD, Oldridge NB. Evaluation of the Health Utilities Index Mark-3 in heart failure. J Card Fail 2011;2:143–50.

197. Marra CA, Rashidi AA, Guh D, Kopec JA, Abrahamowicz M, Esdaile JM, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res 2005;14:1333−44.

198. Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. Reliability of the Health Utilities Index-mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res 1995;4: 249−57.

199. Blanchard C, Feeny D, Mahon JL, Bourne R, Rorabeck C, Stitt L, et al. Is the Health Utilities Index valid in total hip arthroplasty patients? Qual Life Res 2004;13:339−48.

200. Grootendorst P, Feeny D, Furlong W. Health Utilities Index Mark 3: evidence of construct validity for stroke and arthritis in a population health survey. Med Care 2000;38:290−9.

201. Blanchard C, Feeny D, Mahon JL, Bourne R, Rorabeck C, Stitt L, et al. Is the Health Utilities Index responsive in total hip arthroplasty patients? J Clin Epidemiol 2003;56:1046−54.

202. Drummond M. Introducing economic and quality of life measurements into clinical studies. Ann Med 2001;33:344−9.

203. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. Health Serv Res 1976;11:478–507.
204. Kaplan RM, Sieber WJ, Ganiats TG. The quality of well-being scale: comparison of the interviewer-administered version with a self-administered questionnaire. Psych Health 1997;12:783–91.
205. Seiber WJ, Groessl EJ, David KM, Ganiats TG, Kaplan RM. Quality of Well Being Self-Administered (QWB-SA) Scale: user's manual. San Diego: Health Services Research Center, University of California; 2008.
206. Groessl EJ, Kaplan RM, Cronan TA. Quality of well-being in older people with osteoarthritis. Arthritis Rheum 2003;49:23–8.
207. Ganiats TG, Muhlen DG, Kaplan RM, Barrett-Connor E. Gender differences in quality of life in geriatric orthopaedic patients [abstract]. Qual Life Res 1997;6:648.
208. Bombardier C, Raboud J, and the Auranofin Cooperating Group. A comparison of health-related quality-of-life measures for rheumatoid arthritis research. Control Clin Trials 1991;12:243S–56S.
209. Moock J, Kohlmann T. Comparing preference-based quality-of-life measures: results from rehabilitation patients with musculoskeletal, cardiovascular, or psychosomatic disorders. Qual Life Res 2008;17: 485–95.
210. Haywood KL, Garratt AM, Fitzpatrick R. Quality of life in older people: a structured review of generic self-assessed health instruments. Qual Life Res 2005;14:1651–68.
211. Palta M, Chen HY, Kaplan RM, Feeny D, Cherepanov D, Fryback DG. Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. Med Decis Making 2011;2:260–9.
212. Frosch DL, Kaplan RM, Ganiats TG, Groessl EJ, Sieber WJ, Weisman MH. Validity of self-administered quality of well-being scale in musculoskeletal disease. Arthritis Rheum 2004;51:28–33.
213. AQoL instruments. 2009. URL: http://www.aqol.com.au/.
214. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. Qual Life Res 1999;8:209–24.
215. Hawthorne G. Assessing utility where short measures are required: development of the short Assessment of Quality of Life-8 (AQoL-8) instrument. Value Health 2009;12:948–57.
216. Busija L, Buchbinder R, Osborne RH. Quantifying the impact of transient joint symptoms, chronic joint symptoms, and arthritis: a population-based approach. Arthritis Rheum 2009;61:1312–21.
217. Ackerman IN, Graves SE, Wicks IP, Bennell KL, Osborne RH. Severely compromised quality of life in women and those of lower socioeconomic status waiting for joint replacement surgery. Arthritis Rheum 2005;53:653–8.
218. Crotty M, Prendergast J, Battersby MW, Rowett D, Graves SE, Leach G, et al. Self-management and peer support among people with arthritis on a hospital joint replacement waiting list: a randomised controlled trial. Osteoarthritis Cartilage 2009;17:1428–33.
219. Osborne RH, Buchbinder R, Ackerman IN. Can a disease-specific education program augment self-management skills and improve health-related quality of life in people with hip or knee osteoarthritis? BMC Musculoskelet Disord 2006;7:90.
220. Bennell K, Wee E, Coburn S, Green S, Harris A, Staples M, et al. Efficacy of standardised manual therapy and home exercise programme for chronic rotator cuff disease: randomised placebo controlled trial. BMJ 2010;340:c2756.
221. Bennell KL, Hinman RS, Metcalf BR, Buchbinder R, McConnell J, McColl G, et al. Efficacy of physiotherapy management of knee joint osteoarthritis: a randomised, double blind, placebo controlled trial. Ann Rheum Dis 2005;64:906–12.
222. Galea MP, Levinger P, Lythgo N, Cimoli C, Weller R, Tully E, et al. A targeted home- and center-based exercise program for people after total hip replacement: a randomized clinical trial. Arch Phys Med Rehabil 2008;89:1442–7.
223. Buchbinder R, Osborne RH, Ebeling PR, Wark JD, Mitchell P, Wriedt C, et al. A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. N Engl J Med 2009;361:557–68.
224. Hawthorne G. The effect of different methods of collecting data: mail, telephone and filter data collection issues in utility measurement. Qual Life Res 2003;12:1081–8.
225. Hawthorne G, Richardson J, Day NA. Using the Assessment of Quality of Life (AQoL) Instrument, version 1.0. Report no. 12. Melbourne: Centre for Health Program Evaluation, University of Melbourne; 2000.
226. Hawthorne G, Osborne R. Population norms and meaningful differences for the Assessment of Quality of Life (AQoL) measure. Aust N Z J Public Health 2005;29:136–42.
227. Whitfield K, Buchbinder R, Segal L, Osborne RH. Parsimonious and efficient assessment of health-related quality of life in osteoarthritis research: validation of the Assessment of Quality of Life (AQoL) instrument. Health Qual Life Outcomes 2006;4:19.
228. Richardson J, Day NA, Peacock S, Iezzi A. Measurement of the quality of life for economic evaluation and the Assessment of Quality of Life (AQoL) Mark 2 instrument. Aust Health Econ Rev 2004:1;62–88.
229. Osborne RH, Hawthorne G, Lew EA, Gray LC. Quality of life assessment in the community-dwelling elderly: validation of the Assessment of Quality of Life (AQoL) instrument and comparison with the SF-36. J Clin Epidemiol 2003;56:138–47.
230. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. Ann Med 2001;33:358–70.

Summary Table for Adult Quality of Life Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
SF-36	Generic health profile	Self-report Self-administration	Moderate to low	Low	Range 0–100 High scores = better health	Poor–good	Generally good	Good	Availability of norms provides context for score interpretation	Large amount of measurement error for some subscales
SF-12	Generic health profile	Self-report Self-administration	Low	Low	Range 0–100 High scores = better health	Good	Generally good	Fair but need more data	Low respondent burden and availability of norms	Not for monitoring individuals
NHP	Generic health profile	Self-administration	Low	Low	Range 0–100 High scores = worse health	Poor–good	Generally good	Generally good	Relatively simple to complete, low respondent/ administrative burden	Questionable psychometric properties, particularly among people with minor disabilities
SIP	Generic health profile	Self-report Self-administration	High	Moderate	Range 0–100 High scores = worse health	Overall score: good Dimensions: good Subscales: poor to good	Generally good	Generally good, but need more data	Able to detect change among a range of rheumatology interventions	High respondent burden, particularly for SIP136 Poor reliability in some subscales
SF-6D	Health utility measure	Self-report Self-administration	Moderate to low	Moderate to low	Range 0.30–1.00 High scores = better HRQOL	Very good	Very good	Fair	Low respondent burden and availability of norms; scores can be computed whenever SF-36 or SF-12 have been administered	Not suitable for use with populations with severely impaired HRQOL
HUI3	Health utility measure	Self-report Self-administration	High to moderate	High to moderate	Range 0–1 High scores = better HRQOL	Fair–good	Generally good	Fair to good, but need more data	Utility instrument, can be used in cost utility analysis; ability to detect improvement due to rheumatology treatments	Cost Limited sensitivity to deterioration
QWB-SA	Health utility measure	Self-administration	Moderate to low	Low	Range 0–1 High scores = better HRQOL	Insufficient information	Good	Generally good, but need more data	Comprehensive coverage of health state levels with no evidence of floor or ceiling effects	Need further psychometric evaluations in rheumatic conditions
AQoL	Health utility measure	Self-report Self-administration	Very low	Very low	Range −0.04 to 1.00 High scores = better HRQOL	Very good	Very good	Generally good, but need more data	Can be used to make comparisons with the general population	Need further psychometric evaluations in rheumatic conditions

* SF-36 = Medical Outcomes Study Short Form 36; SF-12 = Medical Outcomes Study Short Form 12; NHP = Nottingham Health Profile; SIP = Sickness Impact Profile; SF-6D = Medical Outcomes Study Short Form 6D; HRQOL = health-related quality of life; HUI3 = Health Utilities Index Mark 3; QWB-SA = Quality of Well-Being Scale Self-Administered; AQoL = Assessment of Quality of Life.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S438–S445
DOI 10.1002/acr.20560

Measures of Health Status and Quality of Life in Juvenile Rheumatoid Arthritis

Pediatric Quality of Life Inventory (PedsQL) Rheumatology Module 3.0, Juvenile Arthritis Quality of Life Questionnaire (JAQQ), Paediatric Rheumatology Quality of Life Scale (PRQL), and Childhood Arthritis Health Profile (CAHP)

A. C. CARLE, E. MORGAN DEWITT, AND M. SEID

INTRODUCTION

In this review, we describe four measures of health-related quality of life (HRQOL) designed for children with juvenile rheumatoid arthritis. HRQOL generally refers to how an individual feels about aspects of their life in relation to their health. The World Health Organization originally described HRQOL as minimally including physical, mental, and social health dimensions (1). Subsequent definitions, although varied, have incorporated the fact that individuals have an important and distinct viewpoint regarding their disease and quality of life (2). They have also emphasized the subjective nature of HRQOL (2). These features present unique challenges when measuring HRQOL in children. A child's age and cognitive development may limit their ability to answer and understand questions, requiring proxy-report. Yet research suggests that parents and children do not always view HRQOL similarly and that these differences represent valid differences (3–5). Thus, for each of the measures below, users should evaluate strengths and weaknesses with respect to the perspective(s) they wish to measure and a child's developmental status.

A. C. Carle, PhD, E. Morgan Dewitt, MD, M. Seid, PhD: Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio.

Dr. Carle has received an NIH Co-Principal Investigator PROMIS Supplement (National Institute of Arthritis and Musculoskeletal and Skin Diseases grant 3U01-AR-057940-02S1), and an NIH Principal Investigator National Institute of Nursing Research grant (1R15-NR-010631-01A1). Dr. Dewitt is an NIH PI (Cincinnati Children's Hospital Medical Center Site) with the Patient Reported Outcomes Measurement Information System Cooperative Network (1U01-AR-057940-01).

Address correspondence to A. C. Carle, PhD, 3333 Burnet Avenue, Cincinnati, OH 45226. E-mail: adam.carle@cchmc.org.

Submitted for publication February 4, 2011; accepted in revised form July 9, 2011.

PEDIATRIC QUALITY OF LIFE INVENTORY (PEDSQL) RHEUMATOLOGY MODULE 3.0

Description

Purpose. Varni et al in 1999 (6), designed the PedsQL Generic Core Scales as a generic health-related quality of life (HRQOL) measure for use across the heterogeneous pediatric population, including healthy children and children with diseases. Whereas in 2002, Varni et al (7) developed the PedsQL Rheumatology Module 3.0 to measure pediatric rheumatology-specific HRQOL. The Rheumatology Module measures HRQOL aspects uniquely important to children with rheumatic diseases and complements the core scales. The Rheumatology Module fits within Varni and colleagues' broader efforts to measure HRQOL in pediatric health conditions using the PedsQL Generic Core Scales (6–8).

Content. The 22-item Rheumatology Module measures 5 dimensions: pain-hurt, daily activities, treatment, worry, and communication.

Number of items. 22 items comprise the Rheumatology Module: pain-hurt (4 items), daily activities (5 items), treatment (7 items), worry (3 items), and communication (3 items).

Recall period for items. Respondent's answers address the past month.

Endorsements. None.

Examples of use. Research has used the Rheumatoid Module to examine HRQOL for children with juvenile rheumatoid arthritis (JRA) and children generally (9,10), to investigate coping among children with JRA (11), and explore outcomes (12,13), among other topics.

Practical Application

How to obtain. A copy can be obtained online at www.pedsql.org. The site includes a detailed fee structure description.

Method of administration. The PedsQL Rheumatology Module 3.0 uses parent (proxy) report and child self-report to measure HRQOL. Varni et al (7) report that, when possible, one should measure both parent and child perspectives. Rheumatology Module questions use a 5-point ordinal (i.e., polytomous) scale for a child self-report (ages 8–17 years) and parent proxy-report (ages 2–17 years). Options range from 0 = never a problem to 4 = almost always a problem. Children ages 5–7 years answer using a simplified 3-point scale, with each response anchored to a happy-to-sad-faces scale. A self-report form does not exist for children ages 2–5 years, relying instead on parent proxy-report to measure HRQOL for this age group. Additionally, for children ages 2–5 years, parent proxy-report does not include the worry and communication scales.

Scoring. Items are reverse coded and linearly transformed to a 0–100 scale (e.g., 0 = 100 to 4 = 0). Each scale score equals the average of the transformed items answered in a given scale. For scales with more than 50% missing data, one does not compute a scale score. However, research suggests little missing data occur (7).

Score interpretation. High scores correspond to better quality of life. Cut-scores and minimum clinically important differences have not been established.

Respondent burden. Administration takes approximately 15 minutes for child self-report and 10 minutes for parent proxy-report.

Administrative burden. No data available.

Translations/adaptations. In addition to English, independent research groups have created French, German, Italian, Russian, Slovenian, and Spanish translations. Research has not yet validated these translations (14).

Psychometric Information

Method of development. Varni et al (7) developed the Rheumatology Module using their experience developing previous HRQOL measures, a review of the literature, patient and parent focus groups, item generation, cognitive interviews, pretesting, and field testing of the final instrument in a sample of the target population.

Acceptability. Little missing data on the Rheumatology Module appear to occur (generally $<2\%$) and sufficient proportions of respondents endorse each category.

Reliability. Varni et al (7) examined the reliability and validity of the PedsQL Rheumatology Module 3.0 in a sample of 231 children ages 5–18 years and 244 parents of these (and additional) children ages 5–18 years. Parents and children (ages 8–18 years) self-administered the measures. An interviewer administered the measures to children ages 5–7 years. Cronbach's alpha across scales and forms generally demonstrated acceptable reliability for research, with the majority exceeding 0.70. Several parent proxy-report Cronbach's $\alpha > 0.90$. However, Cronbach's alpha for children ages 5–7 years self-report were generally poor, limiting child self-report for this age range.

Validity. Varni et al (7) demonstrated construct validity by using analysis of variance (ANOVA) to compare groups of children known to differ in the investigated health construct. These analyses found statistically significant differences across several different groups of children with different types of rheumatic diseases (e.g., fibromyalgia versus other rheumatic diseases) for both self-report and proxy-report, supporting construct validity. The authors also established construct validity by examining intercorrelations among the PedsQL total score and the Rheumatology Module scale sores. They found medium to large effect size correlations.

Ability to detect change. The authors (7) demonstrated responsiveness by examining change across time among patients for whom a change was expected. Repeated-measures ANOVAs showed responsiveness for the pain and hurt scale.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PedsQL Rheumatology Module 3.0 constitutes a relatively well-validated measure of multiple dimensions of HRQOL specifically important to children with rheumatic diseases. When accompanied by the PedsQL Generic Core Scales, the two measures reliably and validly cover a broad range of HRQOL dimensions. Varni et al (7) specifically developed the Rheumatology Module to span a very broad age range for child self-report and an even broader age range when including parent-report. Moreover, they accomplished this while maintaining consistent items and scales across forms. This increases the comparability of scores across a wide range of ages and, for a given child, increases comparability across the child's life span.

Caveats and cautions. Some features limit the Rheumatology Module. Research has not used item response theory, structural equation modeling, or confirmatory factor analysis. Without this research, the internal validity of the Rheumatology Module remains unestablished, which limits its interpretability. Finally, the translations have not been examined.

Clinical usability. Research supports usability.

Research usability. Research supports usability.

JUVENILE ARTHRITIS QUALITY OF LIFE QUESTIONNAIRE (JAQQ)

Description

Purpose. Duffy et al (15) developed the JAQQ to measure health-related quality of life (HRQOL) among children with juvenile rheumatoid arthritis (JRA) and juvenile spondylarthritis. They sought to create an easy to use, responsive instrument that measured multiple domains that could uniquely measure areas of importance to individual children.

Content. The JAQQ measures gross motor function, fine motor function, psychosocial function, and general symptoms.

Number of items. The instrument includes a total of 74 items: gross motor function (17 items), fine motor function (16 items), psychosocial function (22 items), and general symptoms (19 items).

Response options/scale. Each item uses a 7-point ordinal scale ranging from 1 (none of the time) to 7 (all of the time). The JAQQ also includes a measure of pain (100-mm pain visual analog scale).

Recall period for items. No data available.

Endorsements. None.

Examples of use. Several avenues of research have included the JAQQ: research comparing parent and child perceptions of HRQOL (16), studies describing the HRQOL of children with juvenile rheumatoid arthritis (17), and outcome studies (18–20).

Practical Application

How to obtain. Copy of the JAQQ can be obtained from Dr. Ciarán Duffy (E-mail: ciaran.duffy@muhc.mcgill.ca).

Method of administration. Parents and/or children (>9 years) self-administer the JAQQ.

Scoring. Respondents answer all items each time they receive the JAQQ. However, at first administration, patients identify 5 items in each domain with which they have the most difficulty. Each dimension's scores are computed as the unweighted average of the 5 items, at baseline and followup. The total score equals the unweighted average of the dimension scores. Respondents can also volunteer items when completing the JAQQ. These patient-generated items can become part of the dimensional score if they are among the 5 identified items. Duffy et al report that this "ensures . . . patient input is incorporated" (15). Change scores comprise differences between administrations.

Score interpretation. High scores correspond to poorer HRQOL.

Respondent burden. The measure takes ~20 minutes to complete at first administration and 5 minutes on subsequent administrations.

Administrative burden. Scoring takes approximately 5–10 minutes.

Translations/adaptations. English, French, and Dutch versions exist.

Psychometric Information

Method of development. An expert panel generated the initial item set. After translating into French and back translating into English, the authors pretested the English and French versions of the questionnaire by interviewing 10 rheumatology clinic patients (parents and children). Final development occurred among 91 patients from the Montreal Children's Hospital arthritis clinic. This included interviews with parents of 40 of the children. Initially-generated items were classified into dimensions by expert opinion and reduced by expert opinion and cluster analysis. In this phase, a school/cognitive function dimension was deleted. The reduction process resulted in 85 items in the 4 domains.

Acceptability. No data available.

Reliability. Using a sample of 369 English children, Shaw et al (17) reported the following Cronbach's alpha values: 0.94 for the gross motor domain, 0.97 for the fine motor domain, 0.93 for the psychosocial domain, 0.88 for the general domain, and 0.96 for the entire scale. To more validly estimate reliability, the authors computed these coefficients based on children's responses to all of the items in the JAQQ (rather than the individualized subset of the most problematic items).

Validity. Pretesting and validation used a sample of 30 patients from the same clinic. To establish construct validity, Duffy et al (15) examined the correlation of the JAQQ dimension and total score with measures of joint disease activity and pain. The authors found moderate correlations between the JAQQ and measure of joint disease activity, with the highest correlations occurring between the JAQQ total score and the fine motor dimension with the sum of joint severity score (r = 0.35 and 0.36, respectively). JAQQ scores correlated relatively well with pain scores, while correlations for the psychosocial dimension were low to moderate with diseases activity (r = 0.19) and pain (r = 0.34). The authors observed mixed correlations for the general symptoms dimension with other scores. These correlations corresponded to the authors' a priori hypotheses, indicating construct validity. With respect to face and content validity, 95% of the 20 experts agreed that the JAQQ addressed the dimensions it claims to measure and more than 80% accepted each of the individual items.

Ability to detect change. To determine responsiveness, the authors compared correlations between JAQQ change scores and change scores on other included measures based on a priori predictions. They found that these correlations generally corresponded to the construct validity pattern (e.g., best between mean JAQQ and pain). Additionally, they indirectly demonstrated responsiveness by showing that the JAQQ discriminated among patients using physician-based global health categorizations. In other work (published as abstracts), Duffy and colleagues have further established the ability of the JAQQ to detect change (21,22). Research has not established cut points or minimum clinically important differences for the JAQQ, nor do normative data exist.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The JAQQ offers a rheumatology-specific measure that incorporates a range of items relevant to a child's physical and psychosocial health and functional status. Duffy et al (15) report that the JAQQ presents a clinical advantage over other measures because it offers individualized assessment. Each child selects the 5 most problematic items in each domain and only these items are scored on the initial and subsequent administrations. This potentially increases the instrument's sensitivity to clinical change and may make it especially useful in clinical settings focused on an individual child. However, this essentially renders the instrument unusable in research. In essence, no 2 children complete the same measure making comparisons across children impossible. It also limits the JAQQ's discriminant validity. Duffy et al (15) have reported that the unique scoring system makes the JAQQ especially suited to clinical trials. However, it is not clear that this is an advantage because the meaning of a change score differs across children, obscuring results describing average change (see Crocker and Algina [23], McDonald [24], or Nunnally and Berstein [25] for discussions of the psychometric properties that scores should have to make them useful in research.)

Caveats and cautions. In pretesting, the authors identified items patients rarely or never endorsed, as well as items that appeared to measure similar things as other items on the 85-item questionnaire. As a result, they trimmed an additional 11 items, resulting in a total of 74 items in 4 dimensions (described above). Thus, while the current form of the JAQQ has 74 items, the validity data correspond to the 85-item version, warranting some caution regarding the validity of the present version. As another limit, published research has not used item response theory, structural equation modeling, or confirmatory factor analysis to evaluate the psychometric properties of the JAQQ. This transpires partly because of the unique method by which patients and parents complete the measure. Without this research, the internal validity and measurement structure of the JAQQ remains unclear, which limits the scales' interpretability. Perhaps future research will make use of item response theory and develop a computerized adaptive test (26) version of the JAQQ, which would simultaneously offer an assessment tailored to a child while still delivering a score comparable across children. The JAQQ does not include a specific dimension to measure HRQOL with respect to school and cognitive ability, which limits the JAQQ's coverage of important HRQOL dimensions in childhood. Finally, research has not investigated the translations.

Clinical usability. The unique scoring system of the JAQQ may make it especially useful in clinical work. By including patient-generated items, the JAQQ should capture important HRQOL issues.

Research usability. Currently unresolved key issues (e.g., reliability and a score's meaning across children) limit its application in research.

PAEDIATRIC RHEUMATOLOGY QUALITY OF LIFE SCALE (PRQL)

Description

Purpose. Believing that the length of existing pediatric health-related quality of life (HRQOL) measures limits their use in clinical care, Filocamo et al (27) sought to develop and validate a short HRQOL measure specific to pediatric rheumatic disease.

Content. The PRQL measures physical health (PhH) and psychosocial health (PsH).

Number of items. The PRQL comprises 10 items total, 5 for each subscale.

Response options/scale. Both parent and child forms use a 4-point ordinal scale (0 = never to 3 = all the time) to measure the frequency of symptoms in the previous month for all items.

Recall period for items. Respondents apply their answers to the previous month.

Endorsements. None.

Examples of use. Other than the study describing its development, no examples of the PRQL's use yet exist.

Practical Application

How to obtain. One can obtain a copy of the parent and child English versions of the instrument by downloading the supplemental material accompanying the article that describes the scale's development (27).

Method of administration. The PRQL has parent proxy-report and child self-report forms.

Scoring. The PhH and PsH scores constitute the total sum of the item responses for each subscale respectively or the total sum for items within each subscale (with specific instructions for scoring items marked not applicable). The total score ranges 0–30, and separate subscale (PhH and PsH) scores each range 0–15. The authors instruct users not to create a total score if more than 2 questions are marked inapplicable in a given scale. The PhH and PsH scores constitute the total sum of the item responses for each subscale respectively or the total sum for items within each subscale (with specific instructions for scoring items marked not applicable).

Score interpretation. High scores correspond to poorer functioning.

Respondent burden. Completion takes ~5 minutes or less.

Administrative burden. Scoring takes ~5 minutes or less.

Translations/adaptations. The PRQL has both Italian and English versions. However, research has not examined the psychometric properties of the English translation of the PRQL.

Psychometric Information

Method of development. A panel of 6 pediatric rheumatologists developed the PRQL. The panel initially identified and derived 389 items through a review of the literature and existing pediatric HRQOL measures, discussion, and semistructured face-to-face interviews with 37 children with pediatric rheumatic diseases and their parents. Subsequently, the panel kept 25 items relevant to the 2 desired domains, general to all pediatric rheumatic diseases, applicable to children of all ages, which expressed a single idea, and about which the entire panel agreed the questionnaire should include. The developers then asked another expert panel (that included pediatric rheumatologists and others) and a convenience sample of 42 children and their parents to comment on and criticize the draft measure. This resulted in the deletion of 15 additional items, ending at the final 10-item measure.

Acceptability. No data available.

Reliability. Using a predominantly female sample (77%) of 472 children with juvenile rheumatoid arthritis, the authors evaluated the psychometric properties of the Italian PRQL. To assess reliability, Filocamo et al (27) had 35 parents complete the PRQL a second time within 24 hours. This resulted in test–retest reliability coefficients of 0.91 for the total score, 0.85 for the PhH subscale, and 0.92 for the PsH subscale. These values support the use of the total score and PsH subscale for use in individual patient analyses and the PhH scale in research (25).

Validity. As part of the validation process, the authors report using exploratory factor analysis, with an orthogonal rotation (that forces all underlying factors to be uncorrelated) to examine the construct validity and internal structure of the PRQL. Internal validity refers to the extent to which data support the hypothesis that the question sets do indeed measure 2 separate constructs. These results indicated a 2-factor solution, providing support for creating 2 subscales. Subsequently, the authors evaluated construct validity by examining the extent to which the PRQL correlated with the Juvenile Arthritis Functioning Scale (JFAS), parent's and/or patient's global assessment of the child's well-being, and pain ratings. The authors predicted and generally observed moderate to high correlations in the expected direction for the PhH subscale with parent's assessment of both child's overall well-being, pain intensity, JAFS score, and tender and active joint counts. The remaining correlation for the PhH subscale and all correlations for the PsH subscale were poor.

In addition to construct validity, Filcamo et al assessed discriminative validity. They did this by examining whether differences in the median total PRQL scores corresponded in the expected direction to physicians' ratings of disease course, changes in disease outcome from previous visit, and assessment of morning stiffness. They also examined whether the proportion of children with a score of 0 (i.e., good HRQOL) corresponded to theoretical expectations across these groups. These results generally supported the PRQL's ability to discriminate (e.g., patients with >30 minutes of morning stiffness had the highest median scores). The authors also demonstrated responsiveness by examining whether patients and their parents rated HRQOL worse than healthy children rated their HRQOL. These results showed that only the PhH scale differentiated between these groups.

Finally, as part of the development process, the authors established face and content validity by consulting a panel of experts (which included pediatric rheumatologists). The entire panel indicated their support of the measure's face validity and the appropriateness and coverage of the measure's content. The authors also established face and content validity by asking a convenience sample of 42 children and their parents to complete and criticize the draft PRQL.

Ability to detect change. The authors used the standardized response mean (SRM; the mean change score across children divided by the SD of the change scores) to evaluate responsiveness to clinical change using changes in parents' and patients' scores at a followup administration 3–9 months after baseline. The parent, patient, and physician ratings of disease course provided external criteria for the SRM. For patients rated as improved by a physician, the total score and both subscales were moderately responsive, however for patients rated as worsened, the total score and PhH subscale demonstrated small responsiveness and poor responsiveness for the PsH subscale. Filcamo et al also identified minimum clinically important differences (MCID) for the parent report. They computed MCID as the average change score that corresponded to a rating by the parent, patient, or physician as slightly improved or slightly worsened from the previous visit. MCID ranged from −1.7 (slightly improved) to 1.5 (slightly worsened) for the total score. However, the confidence intervals (CIs) for these overlapped the score for children with stable disease course. This problem was particularly pronounced for the subscale MCIDs, with the CIs for slightly improved and worsened overlapping even with each other. This indicates that more work is needed to establish MCIDs that discriminate well. Research has not established cut points or normative data.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PRQL delivers a very short measure of 2 dimensions of HRQOL relevant to children with rheumatic diseases. Although its brevity can be a strength (because patients can complete it quickly and clinicians can score it quickly), its brevity means that it does not cover the range of potentially important dimensions and aspects of HRQOL.

Caveats and cautions. Like other measures of HRQOL specific to pediatric rheumatology, published research has not yet applied item response theory, structural equation modeling, or confirmatory factor analysis or other latent variable methods to evaluate the internal validity of the PRQL. Although the authors report conducting exploratory factor analyses, it is unclear whether they used an appropriate analytic technique. They report using factor analysis, but provide a reference for principal components analysis. Factor analyses would more validly have examined the question of interest (24). In addition, it is not clear whether they incorporated the ordered-categorical nature of the data in their model. Research shows that this can lead to spurious dimensions (28) and subsequently biased loading estimates. This limits the interpretability of the published results. Finally, while initial evidence seems to support the validity of the total and PhH subscale scores, the results did not strongly support the PsH subscale score. The psychometric properties of the English translation have not been examined.

Clinical usability. The scale's brevity may make the total score and PhH scores potentially attractive in clinical settings.

Research usability. Several issues (see Caveats and cautions above) limit the PRQL's use in research.

CHILDHOOD ARTHRITIS HEALTH PROFILE (CAHP)

Description

Purpose. Tucker et al (29) developed the CAHP to capture a broad range of health statuses in children with juvenile rheumatoid arthritis (JRA).

Content. The CAHP measures physical functioning, psychosocial functioning, and the disease's effect on the family. It was developed and is intended to be used with the Childhood Health Questionnaire (CHQ). The CAHP includes 3 modules: generic health status (measured by the CHQ), juvenile rheumatoid arthritis–specific scales, and patient characteristics.

Number of items. No data available.

Response options/scale. No data available.

Recall period for items. No data available.

Endorsements. None.

Examples of use. Other than manuscripts discussing measures for measuring health-related quality of life (HRQOL) among children with JRA, we found no examples of the CAHP's use.

Practical Application

How to obtain. No data available.

Method of administration. Parents or teens (age ≥13 years) self administer the CAHP.

Scoring. No data available.

Score interpretation. No data available.

Respondent burden. No data available.

Administrative burden. No data available.

Translations/adaptations. No data available.

Psychometric Information

Method of development. The self-administered instrument was developed using prospective data from 80 children with JRA ages 5–15 years old. A multidisciplinary team that included a pediatric rheumatologist, physiotherapist, nurse, social worker, and a parent of a child with JRA generated the initial parent report CAHP items (30). To date, only an abstract (29) and secondary sources describe the CAHP (30–32).

Acceptability. No data available.

Reliability. Tucker et al (29) report reliability coefficients ranged from 0.84–0.97, supporting reliability.

Validity. Tucker et al (29), focusing on the functional status scales, used factor analysis and multitrait analysis to determine the internal consistency and discriminant validity of the parent-report CAHP. Factor analyses identified 3 latent variables labeled gross motor function, fine motor function, and usual role activities, and the authors used the factor analysis results to assign items to 3 scales measuring these variables. Additional analyses indicated that 96% of the items had higher correlations with their assigned scales than with other scales, supporting discriminant validity. Finally, the specific functional status scales correlated 0.73 with the CHQ's generic physical functioning scale, indicating that the CAHP may measure aspects not captured by generic scales.

Ability to detect change. No data available.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Too little data exist to identify strengths.

Caveats and cautions. Unfortunately, little detailed published work describes the psychometric properties of the CAHP or the methods by which previously reported psychometric properties were obtained. Research has not described the CAHP's response options, recall period, total number of items, scoring method, development or psychometric properties of the teen-report version, psychometric properties of the CAHP's other scales, or other important features of the CAHP. Additionally, it is not clear how one obtains a copy of the CAHP. These features limit its clinical and research utility.

Clinical usability. Too little data exist to evaluate clinical usability.

Research usability. Too little data exist to evaluate research usability.

DISCUSSION

Summarily, investigators have developed a variety of HRQOL measures designed for assessing HRQOL in juvenile rheumatoid arthritis. Hopefully, future research will address the psychometric properties and internal validity of these measures using structural equation modeling and item response theory, as well the relative utility of a disease-specific approach versus a more general approach (e.g., NIH's Patient Reported Outcomes Measurement Information System: PROMIS).

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. World Health Organization. Constitution of the World Health Organization basic document. Geneva: World Health Organization; 1948.
2. Eiser C, Morse R. A review of measures of quality of life for children with chronic illness. Arch Dis Chil;84:205–11.
3. Eiser C, Kopel S. Children's perceptions of health and illness. In: Petrie KJ, Weinman JA, editors. Perceptions of health and illness: current research and applications. Amsterdam: Harwood Academic; 1997. p. 47–76.
4. Seid M, Opipari L, Huang B, Brunner HI, Lovell DJ. Disease control and health-related quality of life in juvenile idiopathic arthritis. Arthritis Rheum 2009;61:393–9.
5. Varni J, Katz E, Colegrove Jr R, Dolgin M. Adjustment of children with newly diagnosed cancer: Cross-informant variance. J Psychosoc Oncol 1995;13:23–38.
6. Varni J, Seid M, Rode C. The PedsQL: measurement model for the pediatric quality of life inventory. Med Care 1999;37:126–39.
7. Varni JW, Seid M, Smith Knight T, Burwinkle T, Brown J, Szer IS. The PedsQL in pediatric rheumatology: reliability, validity, and responsiveness of the Pediatric Quality of Life Inventory Generic Core Scales and Rheumatology Module. Arthritis Rheum 2002;46:714–25.
8. Varni J, Seid M, Kurtin P. PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 Generic Core Scales in healthy and patient populations. Med Care 2001;39:800–12.
9. Varni JW, Limbers CA, Burwinkle TM. Impaired health-related quality of life in children and adolescents with chronic conditions: a comparative analysis of 10 disease clusters an disease categories/severities utilizing the PedsQL 4.0 Generic Core Scales. Health Qual Life Outcomes 2007;5:43.
10. Schwimmer JB, Burwinkle TM, Varni JW. Health-related quality of life of severely obese children and adolescents. JAMA 2003;289:1813–9.
11. Sawyer MG, Whitham J, Roberton DM, Taplin J, Varni J, Baghurst PA. The relationship between health-related quality of life, pain and coping strategies in juvenile idiopathic arthritis. Rheumatology (Oxford) 2004; 43:325–30.
12. Powell M, Seid M, Szer IS. Efficacy of custom foot orthotics in improving pain and functional status in children with juvenile idiopathic arthritis: a randomized trial. J Rheumatol 2005;32:943–50.
13. Brunner HI, Taylor J, Britto MT, Corcoran MS, Kramer SL, Melson PG, et al. Differences in disease outcomes between medicaid and privately insured children: possible health disparities in juvenile rheumatoid arthritis. Arthritis Rheum 2006;55:378–84.
14. Varni J. PedsQL. URL: www.pedsql.org. 2011.
15. Duffy C, Arsenault L, Duffy K, Paquin J, Strawczynski H. The Juvenile Arthritis Quality of Life Questionnaire: development of a new responsive index for juvenile rheumatoid arthritis and juvenile spondyloarthritides. J Rheumatol 1997;24:738–46.
16. April KT, Feldman DE, Platt RW, Duffy CM. Comparison between children with juvenile idiopathic arthritis (JIA) and their parents concerning perceived quality of life. Qual Life Res 2006;15:655–61.

17. Shaw KL, Southwood TR, Duffy CM, McDonagh JE, on behalf of the British Society of Paediatric and Adolescent Rheumatology, Children's Chronic Arthritis Association, Lady Hoare Trust for Physically Disabled Children, and Arthritis Care. Health-related quality of life in adolescents with juvenile idiopathic arthritis. Arthritis Rheum 2006; 55:199–207.
18. Feldman DE, De Civita M, Dobkin PL, Malleson PN, Meshefedjian G, Duffy CM. Effects of adherence to treatment on short term outcomes in children with juvenile idiopathic arthritis. Arthritis Rheum 2007;57: 905–12.
19. McDonagh J, Southwood T, Shaw K. The impact of a coordinated transitional care programme on adolescents with juvenile idiopathic arthritis. Rheumatology (Oxford) 2007;46:161–8.
20. Takken T, van der Net J, Helders PJ. Do juvenile idiopathic arthritis patients benefit from an exercise program? A pilot study. Arthritis Rheum 2001;45:81–5.
21. Duffy C, Watanabe Duffy K, Gibbon M, Yang H, Platt R. Accuracy of functional outcome measures in defining improvement in juveniile idiopathic arthritis [abstract]. Ann Rheum Dis 2000;59:724–5.
22. Duffy CM, Arsenault L, Duffy KN, Paquin JD, Strawczynski H. Validity and sensitivity to change of the Juvenile Arthritis Quality of Life Questionnaire [abstract]. Arthritis Rheum 1993;36 Suppl 9:S144.
23. Crocker L, Algina J. Introduction to classical and modern test theory. Orlando: Holt, Rinehart and Winston; 1986.
24. McDonald RP. Test theory: A unified treatment. Mahwah (NJ): Erlbaum; 1999.
25. Nunnally JC, Berstein I. Psychometric theory. Volume 2D. New York: McGraw-Hill; 1994.
26. Van Der Linden W, Glas C. Computerized adaptive testing: theory and practice. Norwell (MA): Kluwer Academic; 2000.
27. Filocamo G, Schiappapietra B, Bertamino M, Pistorio A, Ruperto N, Magni-Manzoni S, et al. A new short and simple health-related quality of life measurement for paediatric rheumatic diseases: initial validation in juvenile idiopathic arthritis. Rheumatology (Oxford) 2010;49: 1272–80.
28. Bernstein IH, Teng G. Factoring items and factoring scales are different: evidence for multidimensionality due to item categorization. Psychol Bull 1989;105:465–77.
29. Tucker LB, DeNardo BA, Abetz LN, Landgraf JM, Schaller JG. The Childhood Arthritis Health Profile (CAHP): validity and reliability of the condition-specific scales. Arthritis Rheum 1995;38:S183.
30. Duffy C, Tucker L, Burgos-Vargas R. Update on functional assessment tools. J Rheumatol 2000;58 Suppl:11–4.
31. Duffy C, Watanabe Duffy K. Health assessment in the rheumatic diseases of childhood. Curr Opin Rheumatol 1997;9:440–7.
32. Duffy C. Assessing Outcome in Juvenile Idiopathic Arthritis. J Canadian Rheum Assoc 2000;10:4–7.

Summary Table for Pediatric HRQOL Measures in Juvenile Rheumatoid Arthritis*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
PedsQL	Measure broad range of rheumatology specific HRQOL	Parent proxy-report (ages 2–17 years) and child self-report (ages 5–17 years)	20 minutes	Hand scored	Scores range 0–100; high score indicates better HRQOL	Internal consistency good for most forms and excellent for some	Face, content, and construct validity established	Not established	Sound psychometric properties; appropriate for research use	Internal validity not established
JAQQ	Measure multiple rheumatology specific HRQOL domains and uniquely measure areas of importance to individual children	Parent proxy-report and child self-report for children ≥9 years	20 minutes	Hand scored	Higher scores indicate poorer HRQOL	Internal consistency good	Face, content, and construct validity established	Not established	Several sound psychometric properties; useful for incorporating individualized HRQOL measurement	Scoring system limits interpretability and use; internal validity not established
PRQL	Briefly measure 2 HRQOL domains, PhH, PsH	Parent proxy-report (ages 2–17 years) and child self-report	<5 minutes	Hand scored <5 minutes	Total score range 0–30; subscales range 0–15; high scores indicate poorer HRQOL	Test–retest excellent for total score and PsH; good for PhH	Face, content, discriminative, and construct validity established	SRM and MCID established (with caveats to MCID utility)	Short, easily, and quickly administered and scored	Limited support for PsH scale
CAHP	Capture a broad range of health statuses in children with JRA	Parent proxy-report and child self-report for children ≥13 years	15 minutes	Unknown	Unknown	Reliability coefficients range 0.84–0.97	Internal and discriminant validity partially established	Unknown	Not applicable	Limited psychometric information; not recommended for clinical or research use

* HRQOL = health-related quality of life; PedsQL = Pediatric Quality of Life Inventory Rheumatology Module 3.0; JAQQ = Juvenile Arthritis Quality of Life Questionnaire; PRQL = Paediatric Rheumatology Quality of Life Scale; PhH = physical health domain; PsH = Psychosocial health domain; SRM = standardized response mean; MCID = minimum clinically important difference; CAHP = Childhood Arthritis Health Profile; JRA = juvenile rheumatoid arthritis.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S208–S228
DOI 10.1002/acr.20632

Measures of Knee Function

International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS)

NATALIE J. COLLINS,[1] DEVYANI MISRA,[2] DAVID T. FELSON,[2] KAY M. CROSSLEY,[1] AND EWA M. ROOS[3]

INTRODUCTION

Patient-reported measures of knee function are important for the comprehensive assessment of rheumatology conditions in both clinical and research contexts. To merit inclusion in this review, measures of knee function were required to be patient reported and assess aspects considered important by adult patients with knee problems such as injury or osteoarthritis (OA). Therefore, measures used in rheumatology, orthopedics, and sports medicine were considered. Dimensions deemed to be important to patients included pain, function, quality of life, and activity level. To identify instruments fulfilling these criteria, we utilized published reviews of knee instruments (1), knee OA instruments (2), and measures for use in patellofemoral arthroplasty (3).

Based on these reviews, as well as extensive searches of more recent literature, we included the following 9 patient-reported outcomes: Activity Rating Scale, International Knee Documentation Committee Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score, Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form, Knee Outcome Survey Activities of Daily Living Scale, Lysholm Knee Scoring Scale, Tegner Activity Scale, Oxford Knee Score, and Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). Although the WOMAC can be applied to the hip and knee, this study contains data only applicable to the knee. Measures assessing activity level are listed separately.

Psychometric data pertaining to the reliability and responsiveness of each patient-reported outcome are shown in Tables 1 and 2. The number of psychometric reports concerning each instrument ranges from 2–27. A higher number of reports indicates a higher degree of certainty in interpretation of the psychometric properties.

Psychometric properties were based on data provided in Tables 1 and 2, and interpreted using standardized guidelines. Internal consistency was considered adequate if Cronbach's alpha was at least 0.7 (4), and test–retest (intrarater) reliability was adequate if the intraclass correlation coefficient was at least 0.8 for groups and 0.9 for individuals (5). Floor and ceiling effects were considered to be absent if no participants scored the bottom or top score, respectively, and acceptable if $<$15% of the cohort scored the bottom or top score, respectively (6,7). We defined content validity as present when there was patient involvement in the development and/or selection of items (7). Measures were deemed to have face validity if the reviewers considered that the items adequately reflected the measured construct, or if studies reported that expert panels had made a similar assessment (8). Construct validity was considered adequate if expected correlations were found with existing measures that assess similar (convergent construct validity) and dissimilar (divergent construct validity) constructs (7). As there is no gold standard measure of patient-reported outcome, criterion validity is not applicable to this review. Effect sizes of $<$0.5

Dr. Collins's work was supported by a National Health and Medical Research Council (Australia) Health Professional Research Training (Post-Doctoral) Fellowship. Dr. Roos's work was supported by the Swedish Medical Research Council.

[1]Natalie J. Collins, PhD, PT, Kay M. Crossley, PhD, PT: The University of Melbourne, Melbourne, Victoria, Australia; [2]Devyani Misra, MD, David T. Felson, MD, MPH: Boston University School of Medicine, Boston, Massachusetts; [3]Ewa M. Roos, PhD, PT: University of Southern Denmark, Odense, Denmark.

Dr. Roos has received speaking fees (less than $10,000) from Biomet.

Address correspondence to Ewa M. Roos, PhD, PT, Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark, Campusvej 55, DK-5230, Odense M, Denmark. E-mail: eroos@health.sdu.dk.

Submitted for publication February 10, 2011; accepted in revised form May 20, 2011.

were considered small, 0.5–0.8 were considered moderate, and >0.8 were considered large (9). In this context, the minimum clinically important difference is the amount of change of a patient-reported outcome that represents a meaningful change to the patient, while the patient-acceptable symptom state is the least abnormal function score at which patients would consider themselves having acceptable function (10).

INTERNATIONAL KNEE DOCUMENTATION COMMITTEE (IKDC) SUBJECTIVE KNEE EVALUATION FORM

Description

Purpose. To detect improvement or deterioration in symptoms, function, and sports activities due to knee impairment (11).

Intended populations/conditions. Patients with a variety of knee conditions, including ligament injuries, meniscal injuries, articular cartilage lesions, and patellofemoral pain (11).

Version. The IKDC was formed in 1987 to develop a standardized international documentation system for knee conditions. The IKDC Standard Knee Evaluation Form, which was designed for knee ligament injuries, was subsequently published in 1993 (12) and revised in 1994 (13). The IKDC Subjective Knee Evaluation Form was developed as a revision of the Standard Knee Evaluation Form in 1997. It has undergone subsequent minor revisions since its publication in 2001. The items now have the allocated scores next to each possible response. The minimum score for each item has also been changed so that it is now 0, not 1. The scoring of the numerical rating scales for items 2 and 3 has been reversed so that 0 represents the highest level of symptoms and 10 represents the lowest level of symptoms, which is in line with the scoring of the rest of the items.

Content. Three domains: 1) symptoms, including pain, stiffness, swelling, locking/catching, and giving way; 2) sports and daily activities; and 3) current knee function and knee function prior to knee injury (not included in the total score) (11).

Number of items. 18 (7 items for symptoms, 1 item for sport participation, 9 items for daily activities, and 1 item for current knee function).

Response options/scale. Response options vary for each item. Item 6 dichotomizes response into yes/no; items 1, 4, 5, 7, 8, and 9 use 5-point Likert scales; and items 2, 3, and 10 use 11-point numerical rating scales.

Recall period for items. Not specified for items 1, 3, 5, 7, 8, and 9; 4 weeks for items 2, 4, and 6. Function prior to knee injury for item 10a and current function for 10b.

Endorsements. International Cartilage Repair Society; European Society of Sports Traumatology, Knee Surgery, and Arthroscopy; and American Orthopaedic Society for Sports Medicine (AOSSM).

Examples of use. Conditions: knee ligament injury (anterior cruciate ligament [ACL], posterior cruciate ligament [PCL], lateral collateral ligament [LCL], medial patellofemoral ligament), meniscal tears, knee cartilage lesions, osteochondritis dissecans, and traumatic knee dislocation. Interventions: ligament reconstruction (ACL, PCL, LCL, medial patellofemoral ligament), meniscal repair, meniscectomy, microfracture, osteochondral autografts, platelet-rich plasma injections, high tibial osteotomy, and lateral release.

Practical Application

How to obtain. The most recent revision is freely available at the AOSSM web site as part of the IKDC Knee Forms (2000; www.sportsmed.org/tabs/research/ikdc.aspx). Multiple web sites have published versions of the form.

Method of administration. Patient-completed questionnaire. The form has not been validated for administration by interview, either in person or via telephone.

Scoring. The response to each item is scored using an ordinal method (i.e., 0 for responses that represent the highest level of symptoms or lowest level of function). The most recent version has assigned scores for each possible response printed on the questionnaire. Scores for each item are summed to give a total score (excluding item 10a). The total score is calculated as (sum of items)/(maximum possible score) × 100, to give a total score of 100. An online scoring sheet is available (www.sportsmed.org/tabs/research/ikdc.aspx) that provides a patient's raw score and percentile score (relative to age- and sex-based norms). The item regarding knee function prior to knee injury is not included in the total score.

Missing values. The revised scoring method states that, in cases where patients have up to 2 missing values (i.e., responses have been provided for at least 16 items), the total score is calculated as (sum of completed items)/(maximum possible sum of completed items) × 100.

Score interpretation. Possible score range 0–100, where 100 = no limitation with daily or sporting activities and the absence of symptoms.

Normative values. Normative data are available from the general US population, stratified for age, sex, and current/prior knee problems (14).

Respondent burden. 10 minutes to complete (15). It uses simple language that is suitable for patients.

Administrative burden. Approximately 5 minutes to score. Training is not necessary. Manual scoring can be performed easily using the scoring instructions supplied with the questionnaire.

Translations/adaptations. Available in English, traditional Chinese (Taiwan, Hong Kong), simplified Chinese (China, Singapore), French, German, Italian, Japanese, Korean, Portuguese (Brazil), and Spanish. Cross-cultural adaptations have been conducted for the Brazilian (16), Chinese (17), Dutch (18), Italian (15), and Thai (19) translations.

Psychometric Information

Method of development. The initial set of items was developed by the IKDC, considering questions from the Standard Knee Evaluation Form, the MODEMS Lower

Table 1. Summary of reliability data*

	Patient cohort evaluated (ref.)	Internal consistency (Cronbach's α)	Test–retest (ICC)	MDC	SEM
Function measures					
IKDC	Knee injuries (ACL, meniscal, chondral) (15,20,23)	0.77–0.91	0.90–0.95†	8.8–15.6†	3.2–5.6†
	Cohort of mixed knee pathologies (11,16–18,21)	0.92–0.97	0.87–0.99†	6.7	2.4–4.6†
KOOS	Knee injuries (25,27,32,36)	Pain: 0.84–0.91 Symptoms: 0.25–0.75 ADL: 0.94–0.96 Sport/rec: 0.85–0.89 QOL: 0.64–0.9	Pain: 0.85–0.93 Symptoms: 0.83–0.95 ADL: 0.75–0.91 Sport/rec: 0.61–0.89 QOL: 0.83–0.95	Pain: 6–6.1 Symptoms: 5–8.5 ADL: 7–8 Sport/rec: 5.8–12 QOL: 7–7.2	Pain: 2.2 Symptoms: 3.1 ADL: 2.9 Sport/rec: 2.1 QOL: 2.6
	Knee OA (28–31,33)	Pain: 0.65–0.94 Symptoms: 0.56–0.83 ADL: 0.78–0.97 Sport/rec: 0.84–0.98 QOL: 0.71–0.85	Pain: 0.8–0.97 Symptoms: 0.74–0.94 ADL: 0.84–0.94 Sport/rec: 0.65–0.92 QOL: 0.6–0.91	Pain: 13.4 Symptoms: 15.5 ADL: 15.4 Sport/rec: 19.6 QOL: 21.1	Pain: 7.2–10.1 Symptoms: 7.2–9 ADL: 5.2–11.7 Sport/rec: 9–24.6 QOL: 7.4–10.8
KOOS-PS	Knee OA (40–42)	0.89	0.85–0.86	–	–
KOS-ADL	Mixed knee pathologies (43,47,49–52)	0.89–0.98	0.94–0.98	11.4	4.1
Lysholm Knee Scoring Scale	Knee injuries (ACL, meniscal, chondral; patellar dislocation) (54,55,61,63,64)	0.65–0.73	0.88–0.97	8.9–10.1	3.2–3.6
	Mixed knee pathologies (43,47,119,120)	0.60–0.73	0.68–0.95	–	9.7–12.5†
OKS	Knee OA (46,66,71,121)	0.87–0.93	0.91–0.94	6.1	2.2
WOMAC	Chondral defects (23)		Pain: 0.81–0.85 Symptoms: 0.75–0.86 Function: 0.86–0.93	Pain: 14.4–16.2 Symptoms: 22.9–30.6 Function: 10.6–15	Pain: 5.2–5.8 Symptoms: 8.3–11.1 Function: 3.8–5.4
	Knee OA (42,46,91,92, 94–98,100,101,103–105,108,122,123)	Pain: 0.67–0.92 Symptoms: 0.7–0.94 Function: 0.82–0.98	Pain: 0.65–0.98 Symptoms: 0.52–0.89 Function: 0.71–0.96	Pain: 18.8–22.4 Symptoms: 27.1–29.1 Function: 13.1–13.3	Pain: 6.8–8.1 Symptoms: 9.8–10.5 Function: 4.7–4.8
Activity measures					
ARS	Baseline knee athletic activity for cohort of mixed knee pathologies (113)	–	0.97	–	–
TAS	Knee injuries (ACL, meniscal patellar dislocation) (55,61,64)	n/a	0.82–0.92†	1.0	0.4–0.64
	Knee OA (117)	n/a	0.84	–	–

* ICC = intraclass correlation coefficient; MDC = minimal detectable change; SEM = standard error of measurement; IKDC = International Knee Documentation Committee Subjective Knee Evaluation Form; ACL = anterior cruciate ligament; KOOS = Knee Injury and Osteoarthritis Outcome Score; ADL = activities of daily living; sport/rec = sport/recreation; QOL = quality of life; OA = osteoarthritis; KOOS-PS = Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form; KOS-ADL = Knee Outcome Survey Activities of Daily Living Scale; OKS = Oxford Knee Score; WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index; ARS = Activity Rating Scale; TAS = Tegner Activity Scale; n/a = not applicable.
† Large variation in time between test—retest (up to 12 months).

Limb Instrument, and the Activities of Daily Living and Sports Activity Scales of the Knee Outcome Survey. Pilot testing of the initial version (n = 144) resulted in revision or deletion of existing items and the addition of new items. Testing of the second version (n = 222) resulted in further revisions and deletions (based on missing data), producing a final version. Item-response theory was used to create the scoring system. Patients were not involved in development; rather, items were selected by the IKDC, a committee of international orthopedic surgeons (11).

Acceptability. Missing data were relatively common in testing of the final version of the form, with 57 of 590 patients failing to answer >3 items of 18 (11). Studies consistently report no floor or ceiling effects (i.e., no participants scored lowest or highest score) (11,15,16, 18,20).

Reliability. Internal consistency is adequate for patients with knee injuries and mixed knee pathologies (Table 1). Test–retest reliability is adequate for groups of patients with knee injuries and mixed pathologies and individuals with knee injuries. However, test–retest reliability is slightly below adequate for individuals who fall into a broader category of knee pathologies. The minimal detectable change has been reported to be between 8.8 and 15.6, and the standard error of the measure between 3.2 and 5.6.

Table 2. Summary of responsiveness data*

	Patient cohort evaluated	ES	SRM	MCID
Function measures				
IKDC	Knee injuries (ACL, meniscal, chondral) (20,23)	Meniscal repair/resection (12 m): 2.11 Various cartilage procedures: 0.76 (6 m), 1.06 (12 m)	Meniscal repair/resection (12 m): 1.5 Various cartilage procedures: 0.57 (6 m), 1.0 (12 m)	Chondral injuries: 6.3 (6 m), 16.7 (12 m)
	Cohort of mixed knee pathologies (22,24)	Various surgical procedures (6–28 m): 1.13	Various surgical procedures: 4.4 (4–8 m), 0.94 (6–28 m)	6–28 m: 11.5 (sensitive), 20.5 (specific)
KOOS	Knee injuries (25,27,36)	Partial meniscectomy (3 m): 1.11 (pain), 0.93 (symp.), 0.67 (ADL), 0.9 (sport/rec), 1.15 (QOL) ACLR (6 m): 0.84 (pain), 0.87 (symp.), 0.94 (ADL), 1.16 (sport/rec), 1.65 (QOL) ACI, MF (3 y): 0.82 (pain), 0.72 (symp.), 0.7 (ADL), 0.98 (sport/rec), 1.32 (QOL)	ACI, MF (3 y): 0.71 (pain), 0.61 (symp.), 0.75 (ADL), 0.87 (sport/rec), 0.76 (QOL)	–
	Knee OA (28,31,33)	PT (4 w): 1.08 (pain), 0.97 (symp.), 1.07 (ADL), 0.79 (sport/rec), 0.78 (QOL) TKR (3 m): 2.59 (pain), 1.63 (symp.), 2.52 (ADL), 1.31 (sport/rec), 2.8 (QOL) TKR (6 m): 2.28 (pain), 1.24 (symp.), 2.25 (ADL), 1.18 (sport/rec), 2.86 (QOL) TKR (12 m): 2.55 (pain), 1.59 (symp.), 2.56 (ADL), 1.08 (sport/rec), 3.54 (QOL)	PT (4 w): 1.28 (pain), 1.02 (symp.), 1.37 (ADL), 0.83 (sport/rec), 0.87 (QOL) TKR (3 m): 1.85 (pain), 1.45 (symp.), 1.8 (ADL), 0.89 (sport/rec), 1.93 (QOL) TKR (6 m): 1.67 (pain), 0.99 (symp.), 1.7 (ADL), 0.81 (sport/rec), 1.6 (QOL) TKR (12 m): 2.12 (pain), 1.25 (symp.), 1.9 (ADL), 0.88 (sport/rec), 1.99 (QOL)	–
KOOS-PS	Knee OA (40–42)	PT (4 w): 0.5–0.88 HAI (4 w): 0.51	PT (4 w): 0.73–1.21 HAI (4 w): 0.8 TKR (6 m): 1.4	–
KOS-ADL	Mixed knee pathologies (43,45–47)	PT: 0.44 (1 w), 0.94 (4 w), 1.26 (8 w) PT (6 w): 0.63 TKR (6 m): 1.3	PT (6 w): 7.1 TKR (6 m): 1.1	PFPS: 7.1
Lysholm Knee Scoring Scale	Knee injuries (ACL, meniscal, chondral; patellar dislocation) (55,61,63)	ACLR: 1.0 (6–9 m), 1.1 (1–2 y) Meniscal repair (1 y): 1.2 MF (1–6 y): 1.2	ACLR: 0.93 (6 m), 1.1 (9 m), 1.2 (1 y), 0.93 (2 y) Meniscal repair (1 y): 0.97–1.13 MF (1–6 y): 1.1	–
	Mixed knee pathologies (47,62,120)	PT (1 m): 0.9	Variety of nonsurgical and surgical interventions (3 m): 0.9	–
OKS	Knee OA (46,66)	TKR (6 m): 0.9–2.19	TKR (6 m): 0.7	–
WOMAC	Chondral defects (23)	Various cartilage surgeries (6 m): 0.98 (pain), 0.51 (symp.), 0.88 (function) Various cartilage surgeries (12 m): 1.14 (pain), 0.72 (symp.), 1.2 (function)	Various cartilage surgeries (6 m): 0.91 (pain), 0.40 (symp.), 0.86 (function) Various cartilage surgeries (12 m): 0.94 (pain), 0.64 (symp.), 1.13 (function)	–

(continued)

Table 2. *(Cont'd)*

	Patient cohort evaluated	ES	SRM	MCID
	Knee OA (42,46,92, 96,97,100,101,105, 106,108,124–128)	TKR (3 m): 1.62 (pain), 1.26 (symp.), 2.02 (function) TKR (6 m): 0.95–1.9 (pain), 0.88–1.5 (symp.), 1.01–2.2 (function) TKR (1 y): 1.8–2.4 (pain), 1.8–3.1 (function) TKR (2 y): 1.9–41 (pain), 1.3–24 (symp.), 1.7–23.9 (function) Exercise (2 w): 0.74–0.88 (pain), 0.32–0.44 (symp.), 0.50–0.79 (function) Exercise (6 m): 0.41 (pain), 0.28 (function) Rehabilitation (not defined): 0.52 (pain), 0.42 (symp.), 0.44 (function) Drug (2 w): 0.94 (pain), 0.46 (symp.), 0.72 (function) Drug (3 w): 0.76–0.88 (pain), 0.59–0.63 (symp.), 0.75–0.77 (function) Drug (4 w): 0.69 (pain), 0.41 (symp.), 0.56 (function) Drug (6 w): 0.53–0.8 (pain), 0.6–0.75 (symp.), 0.58–0.82 (function) Drug (8 w): 0.58 (pain), 0.53 (symp.), 0.76 (function) Drug (12 w): 0.44–0.91 (pain), 0.55–0.84 (symp.), 0.58–0.81 (function) Acupuncture (3 w): 0.4 (pain), 0.52 (symp.), 0.31 (function) Acupuncture (8 w): 1.3 (pain), 1.2 (function)	TKR (3 m): 1.14–1.58 (pain), 1.15 (symp.), 1.02–2.02 (function) TKR (6 m): 0.95–1.8 (pain), 0.63–1.3 (symp.), 0.9–1.9 (function) TKR (2 y): 1.55 (pain), 1.03 (symp.), 1.32 (function) Drug (2 w): 1.09 (pain), 0.43 (symp.), 0.89 (function) Exercise (2 w): 0.78–1 (pain), 0.29–0.52 (symp.), 0.69–0.94 (function)	NSAIDs (4 w, function): 9.1 (absolute), 26 (relative) TKR (6 m): 22.87 (pain), 14.43 (symp.), 19.01 (function) TKR (12 m): 36 (pain), 33 (function) TKR (2 y): 27.98 (pain), 21.35 (symp.), 20.84 (function)
Activity measures				
ARS	Baseline knee athletic activity for cohort of mixed knee pathologies	–	–	–
TAS	Knee injuries (ACL, meniscal; patellar dislocation) (55,61)	Various meniscal surgeries (12 m): 0.61 (isolated lesions), 0.84 (combined lesions) ACLR: 0.74 (6 m), 1.1 (9 m), 1.0 (1 y), 1.0 (2 y)	Various meniscal surgeries (12 m): 0.6 (isolated lesions), 0.7 (combined lesions) ACLR: 0.61 (6 m), 0.84 (9 m), 0.96 (1 y), 1.0 (2 y)	–
	Knee OA	–	–	–

* ES = effect size; SRM = standardized response mean; MCID = minimum clinically important difference; IKDC = International Knee Documentation Committee Subjective Knee Evaluation Form; ACL = anterior cruciate ligament; KOOS = Knee Injury and Osteoarthritis Outcome Score; symp. = symptoms; ADL = activities of daily living; sport/rec = sport/recreation; QOL = quality of life; ACLR = ACL reconstruction; ACI = autologous chondrocyte implantation; MF = microfracture; OA = osteoarthritis; PT = physical therapy; TKR = total knee replacement; KOOS-PS = Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form; HAI = intraarticular hyaluronic acid injection; KOS-ADL = Knee Outcome Survey Activities of Daily Living Scale; PFPS = patellofemoral pain syndrome; OKS = Oxford Knee Score; WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index; NSAIDs = nonsteroidal antiinflammatory drugs; ARS = Activity Rating Scale; TAS = Tegner Activity Scale.

Validity. *Face and content validity.* The domains covered by the IKDC appear to represent elements that are likely to be important to patients. However, the lack of patient contribution to the selection and revision of items in the IKDC means that content validity cannot necessarily be assumed.

Construct validity. There are consistent reports of high convergent and divergent construct validity, with the IKDC more strongly correlated with the Short Form 36 (SF-36) physical subscales and component summary than with the mental subscales and component summary (11,16–18,20,21). Studies have shown the IKDC score to be highly correlated with the Cincinnati Knee Rating System, pain visual analog scale, Oxford 12 Questionnaire, Western Ontario and McMaster Universities Osteoarthritis Index, Lysholm score, and SF-36 physical component, physical function, and bodily pain subscales (16,18,22).

Ability to detect change. In patients undergoing surgical treatment of meniscal injury, the IKDC shows large effect sizes at 1 year (Table 2). For patients who have had surgical intervention for cartilage injury, the IKDC shows moderate effect sizes at 6 months and large effect sizes at 1 year. Large effect sizes have been reported from 6–28 months following various surgical procedures conducted in a mixed cohort of knee pathologies. The minimum clinically important difference has been reported to be 6.3 at 6 months and 16.7 at 12 months following cartilage repair (23), and 11.5–20.5 (range 6–28 months) in those who have undergone various surgical procedures for mixed (various) knee pathologies (24). The patient-acceptable symptom state has not been determined.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. At face value, the domains covered by the IKDC appear to represent elements that are likely to be important to patients. It shows adequate internal consistency and has no floor or ceiling effects across mixed groups of patients with knee conditions. The IKDC has been shown to be responsive to change following surgical interventions, highlighting its usefulness in this patient population.

Caveats and cautions. Despite demonstrating face validity, the lack of patient contribution to item selection indicates that content validity cannot necessarily be assumed. The relatively long recall period associated with 3 of the items may be a problem for some patients. The use of 1 aggregate score to represent symptoms, activities, and function may mask deficits in 1 domain. Psychometric testing is lacking for patients with knee osteoarthritis as an isolated group, as well as responsiveness following nonsurgical management, highlighting areas for future studies.

Clinical usability. The IKDC involves minimal administrative and respondent burden, and can be easily scored in the clinic using the online scoring sheet. However, clinicians using the online scoring system need to keep in mind that the normative data provided are from a particular population, and may not be representative of their individual patient's population. Test–retest reliability for those with various knee pathologies suggests that the IKDC may demonstrate inadequate reliability for the evaluation of individual patients.

Research usability. Psychometric evaluation supports the use of the IKDC in research for a variety of knee conditions. As some versions of the IKDC published online contain subtle differences in the wording of instructions and items, researchers should ensure that they utilize the version published as a component of the 2000 IKDC Knee Forms to ensure that findings of psychometric properties still apply, and that comparisons can be made with previous studies. Administrative and respondent burden would not limit research use, although researchers should be diligent in checking for missing data.

KNEE INJURY AND OSTEOARTHRITIS OUTCOME SCORE (KOOS)

Description

Purpose. To measure patients' opinions about their knee and associated problems over short- and long-term followup (1 week to decades).

Intended populations/conditions. Young and middle-aged people with posttraumatic osteoarthritis (OA), as well as those with injuries that may lead to posttraumatic OA (e.g., anterior cruciate ligament [ACL], meniscal, or chondral injury) (25).

Version. The original KOOS remains unchanged, although a short form for function has been developed.

Content. Five domains: 1) pain frequency and severity during functional activities; 2) symptoms such as the severity of knee stiffness and the presence of swelling, grinding or clicking, catching, and range of motion restriction; 3) difficulty experienced during activities of daily living (ADL); 4) difficulty experienced with sport and recreational activities; and 5) knee-related quality of life (QOL) (25).

Number of items. 42 items across 5 subscales.

Response options/scale. All items are rated on a 5-point Likert scale (0–4), specific to each item.

Recall period for items. Previous week for pain, symptoms, ADL, and sport/recreation subscales. Not defined for QOL subscale.

Endorsements. International Cartilage Repair Society, American Academy of Orthopedic Surgeons, and US Food and Drug Administration.

Examples of use. Conditions: knee ligament injury (ACL, posterior cruciate ligament [PCL], medial collateral ligament [MCL]), meniscal tears, knee cartilage lesions, knee OA, and osteochondritis dissecans. Interventions: ligament reconstruction (ACL, PCL, MCL), meniscectomy, microfracture, osteochondral autografts, tibial osteotomy, total knee replacement (TKR), exercise (land based, aquatic), intraarticular sodium hyaluronate injection, pharmacologic therapy, and glucosamine supplementation.

Practical Application

How to obtain. The KOOS and associated documentation are freely available at www.koos.nu.

Method of administration. Patient-completed, in-person questionnaire. The KOOS has not been validated for use during an in-person or telephone interview.

Scoring. Scoring sheets (manual and computer spreadsheets) are provided on the web site. Each item is scored from 0–4. The 5 dimensions are scored separately as the sum of all corresponding items. A total score has not been validated and is not recommended. Scores are then transformed to a 0–100 scale (percentage of total possible score achieved), where 0 = extreme knee problems and 100 = no knee problems (25).

Missing values. If a mark is placed outside a box, the closest box is chosen. If 2 boxes are marked, that which indicates more severe problems is chosen. One or 2 missing values within a subscale are substituted with the average value for that subscale. If >2 items are missing, the response is considered invalid and a subscale score is not calculated.

Score interpretation. 0 = extreme problems and 100 = no problems.

Normative values. Population-based normative data are available, stratified by age and sex (26).

Respondent burden. The KOOS takes 10 minutes to complete (25). It uses simple language and similar 1-word responses for each item. The items largely reflect signs and symptoms of their knee condition and how this affects everyday tasks, so it is not considered that they would have an emotional impact on the individual. The knee-related QOL subscale could be considered the most emotionally sensitive component, as it requires the individual to reflect on how their knee affects their QOL.

Administrative burden. Approximately 5 minutes to score, using the scoring spreadsheet. Training is not necessary, as the components of the KOOS and the scoring instructions are self-explanatory.

Translations/adaptations. Available in English and Swedish (original versions developed concurrently), Austria-German, Czech, Chinese, Croatian, Danish, Dutch, Estonian, French, German, Italian, Japanese, Latvian, Lithuanian, Norwegian, Persian, Portuguese, Polish, Russian, Singapore English, Slovak, Slovenian, Spanish (US), Spanish (Peru), Thai, Turkish, and Ukrainian. Cross-cultural adaptations have been conducted for the Swedish (27,28), Chinese (29), Dutch (30), French (31), Persian (32), Portuguese (33), Russian (Golubev; www.koos.nu), Singapore English (29), Thai (34), and Turkish (35) translations.

Psychometric Information

Method of development. Items were selected based on: 1) the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), version 3.0; 2) a literature review; 3) an expert panel (patients referred to physical therapy for knee injuries, orthopedic surgeons, and physical therapists from Sweden and the US); and 4) a pilot study of 2 questionnaires (1 for symptoms of ACL injury, 1 for symptoms of OA) in individuals with posttraumatic OA. Item-response theory was not used in the development of KOOS or for item selection (25).

Acceptability. Reported rates of missing data are low: 0.8% of items in patients who have undergone knee arthroscopy (27) and 3.2% of items on the pain, symptoms, ADL, and QOL subscales in patients prior to TKR (28). However, patients scheduled for TKR have also exhibited high rates of "not applicable" or missing items (74%) on the sport/recreation subscale (28). Studies consistently report no or acceptable floor or ceiling effects in knee injury cohorts (27,32,36) and in patients with mild or moderate knee OA (28,29,31,33). In those with severe OA awaiting TKR (28–31,33), there are consistent reports of floor effects for the sport/recreation subscale (16–73.3% scored lowest score), and ceiling effects have been reported for the pain (15–22%), sport/recreation (16%), and QOL (17%) subscales up to 12 months following TKR (28).

Reliability. For patients with knee injuries, the pain, ADL, and sport/recreation subscales have adequate internal consistency in all reports, while the symptom and QOL subscales have had reports of lower as well as adequate internal consistency (Table 1). In patients with knee OA, the ADL, sport/recreation, and QOL subscales have adequate internal consistency, while the pain and symptoms subscales have reports of lower as well as adequate internal consistency. Test–retest reliability is adequate for group evaluation in all reports on the pain, symptoms, and QOL subscales for patients with knee injuries, while there are reports of lower and adequate reliability, respectively, for the ADL and sport/recreation subscales. In knee OA, pain and ADL subscales have adequate test–retest data, while for the other subscales, reports indicate both lower and adequate test–retest reliability. Across the 5 subscales, the minimal detectable change ranges from 6–12 for knee injuries and from 13.4–21.1 for knee OA. The standard error of the measure is reported to be lower for knee injuries than for OA.

Validity. *Face and content validity.* As well as exhibiting face validity, the direct involvement of patients with knee conditions in the development of the KOOS facilitates content validity (25,28).

Construct validity. Multiple studies report that the KOOS demonstrates convergent and divergent construct validity, with the KOOS more strongly correlated with subscales of the Short Form 36 (SF-36) that measure similar constructs (e.g., ADL with physical function, sport/recreation with physical function, pain with bodily pain), and less strongly with SF-36 subscales that measure mental health (25,27–30,32,33,36,37). Rasch analysis conducted using patient data 20 weeks post–ACL reconstruction showed that only the sport/recreation and QOL subscales exhibited unidimensionality, not the 3 subscales that were based on the WOMAC (38). A more recent study reported that the KOOS subscales had acceptable dimensionality (37).

Ability to detect change. The KOOS appears to be responsive to change in patients with a variety of conditions that have been treated with nonsurgical and surgical interventions (Table 2). In patients who have undergone partial meniscectomy 3 months previously, large effect sizes are seen on all but the ADL subscale. Large effect sizes are seen in all subscales 6 months after ACL reconstruction. Three years following autologous chondrocyte implantation or microfracture, large effect sizes are seen for the pain, sport/recreation, and QOL subscales, and moderate effects on the symptoms and ADL subscales. In those with knee OA who have undergone physical therapy treatment, large effect sizes are seen at 4 weeks on the pain, symptoms, and ADL subscales, while the sport/recreation and QOL subscales show moderate effects. Large effect sizes are consistently reported on all subscales 3–12 months after TKR. The minimum clinically important difference (MCID) and patient-acceptable symptom state (PASS) have not been calculated in any patient population.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The KOOS has undergone a substantial amount of psychometric testing, largely among populations for whom the scale was intended. Establishment of the KOOS as a reliable and valid measure across multiple languages highlights its usefulness as a patient-reported measure of knee function for people with knee OA and various combinations of ACL, meniscal, and cartilage injury. The use of individual scores for each subscale, rather than an aggregate score, enhances clinical interpretation and in research acknowledges the impact of different interventions on different dimensions (e.g., exercise therapy is likely to have more impact on ADL and sport/recreation, while pharmacology may impact more on pain and symptoms) and ensures content validity in groups of different ages and functional activity levels (e.g., the sport/recreation subscale is more important in patients with a high physical activity level, while the ADL subscale is more important in subjects with a lower physical activity level).

Caveats and cautions. The KOOS has not been validated for interview administration, meaning that it may not be appropriate for patients who are unable to read or write, or where telephone followup is necessary. Rasch analysis suggests that only the subscales that are not based on the WOMAC exhibit unidimensionality in patients who have undergone ACL reconstruction. When administering the KOOS in older or less physically active individuals, higher level components of the ADL and sport/recreation subscales may not be applicable, and could result in missing data. It may be appropriate to leave out the sport/recreation subscale in those with more advanced disease or disability; however, doing so omits the ability to measure improvements seen in these more demanding functions following treatment (28). The MCID and PASS are lacking from psychometric evaluation.

Clinical usability. The KOOS is freely available online. Administration and scoring burden are minimal when online score sheets are utilized. Clinicians should bear in mind that the sport/recreation subscale may not be applicable for less physically active patients, and may not have adequate test–retest reliability in individuals with knee injuries.

Research usability. The KOOS fulfills desired criteria for research outcomes, demonstrating adequate reliability for use in groups and validity when used in those with knee injuries and knee OA. The inclusion of the 3 WOMAC subscales facilitates comparison of findings with studies that have utilized the WOMAC as a primary measure. The lack of reported MCID in any knee condition is a weakness.

KNEE INJURY AND OSTEOARTHRITIS OUTCOME SCORE PHYSICAL FUNCTION SHORT FORM (KOOS-PS)

Description

Purpose. Patients' opinions about the difficulties they experience with physical activity due to their knee problems.

Intended populations/conditions. Knee osteoarthritis (OA).

Version. No modifications since the original publication (39).

Content. Measure of physical function derived from the activities of daily living and sport/recreation subscales of the KOOS (39). Patients rate the degree of difficulty they have experienced over the previous week due to their knee pain, with respect to: 1) rising from bed, 2) putting on socks/stockings, 3) rising from sitting, 4) bending to the floor, 5) twisting/pivoting on injured knee, 6) kneeling, and 7) squatting.

Number of items. 7 items.

Response options/scale. All items are scored on a 5-point Likert scale (none, mild, moderate, severe, extreme) scored from 0–4.

Recall period for items. Previous week.

Endorsements. Osteoarthritis Research Society International and Outcome Measures in Rheumatology Clinical Trials.

Examples of use. Conditions: knee OA. Interventions: total knee replacement (TKR), intraarticular hyaluronic acid injection, and physical therapy.

Practical Application

How to obtain. The KOOS-PS and associated documentation are freely available at www.koos.nu.

Method of administration. Patient-completed questionnaire. Has not been validated for use during in-person or telephone interview.

Scoring. Each question is scored from 0–4. The raw score is the sum of the 7 items. The interval score from 0–100 is obtained using a conversion chart (39).

Missing values. No instructions on how to handle missing values.

Score interpretation. Possible raw score range: 0–28. Scores are then transformed to a score from 0–100, where 0 = no difficulty.

Normative values. Not available.

Respondent burden. Based on findings for the KOOS, no more than 2 minutes to complete. Uses simple language and the same 1-word responses for each of the 7 items. As the items relate to everyday tasks, it is not considered that they would have an emotional impact on the individual.

Administrative burden. Less than 5 minutes to score, using the conversion table provided (39). Training is not necessary, as the questionnaire and scoring instructions are self-explanatory.

Translations/adaptations. Available in English, Swedish, French, and Portuguese. Can easily be compiled by extracting the 7 items needed from the full KOOS forms in all languages in which the KOOS is available. Cross-cultural adaptations have been conducted for the French (40) and Portuguese (41) translations.

Psychometric Information

Method of development. Rasch analysis was conducted on KOOS and Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) data from individuals with knee OA from Sweden, Canada, France, Estonia, and The Netherlands. Patient data from 13 data sets were used (age 26–95 years, male:female ratio 1:1.4). This included community and clinical samples, such as those who had undergone previous meniscectomy, tibial osteotomy, or anterior cruciate ligament repair, as well as those scheduled to undergo TKR (39).

Acceptability. Rates of missing data have not been reported. Findings of 1 study indicate no floor or ceiling effects when used in patients with knee OA (i.e., no patients had lowest or highest score, respectively) (40).

Reliability. The KOOS-PS has adequate internal consistency and test–retest reliability for groups of patients with knee OA; however, its reliability is lower than adequate for use in individuals with knee OA (Table 1). The minimal detectable change and standard error of the measure have not been reported.

Validity. *Face and content validity.* As items are taken directly from the KOOS, which has face and content validity, this can also be assumed for the KOOS-PS.

Construct validity. The KOOS-PS shows evidence of convergent and divergent construct validity. Higher correlations have been shown with the Short Form 36 (SF-36) physical function, role physical, and bodily pain subscales; WOMAC function subscale (excluding KOOS-PS items); and Osteoarthritis Knee and Hip Quality of Life questionnaire (OAKHQOL) physical activity domain (40–42). Conversely, lower correlations have been reported with KOOS pain, symptoms, and quality of life subscales; SF-36 mental health subscales; mental health questionnaires (e.g., Profile of Mood States, Hospital Anxiety and Depression Scale); and OAKHQOL social support (40–42).

Ability to detect change. In patients with knee OA, the KOOS-PS shows moderate to large effect sizes following 4 weeks of physical therapy, and moderate effects 4 weeks after intraarticular hyaluronic acid injection (Table 2). The KOOS-PS is also able to discriminate groups of patients based on use of walking aids (41). The minimum clinically important difference (MCID) and patient-acceptable symptom state have not been reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The KOOS-PS is one of the few knee-related patient-reported outcomes that utilized Rasch analysis in its development. Its inclusion of only 7 items facilitates use with short measures of other dimensions, such as pain visual analog scales, and makes it ideal for those for which long questionnaires may be onerous (e.g., older populations).

Caveats and cautions. The KOOS-PS was intended for use in those with knee OA, and has only undergone psychometric testing for this patient group. The MCID has not been reported.

Clinical usability. The minimal administration and scoring burden associated with the KOOS-PS make it ideal for clinical use, particularly considering that the included items are frequently asked in the standard clinical examination. However, clinicians should bear in mind that the reliability has been shown to be less than adequate for individuals.

Research usability. Psychometric testing shows the KOOS-PS to be valid and reliable for use in groups with knee OA, making it an ideal tool for measuring knee-related function in research.

KNEE OUTCOME SURVEY ACTIVITIES OF DAILY LIVING SCALE (KOS-ADL)

Description

Purpose. To determine symptoms and functional limitation in usual daily activities caused by various knee pathologies (43).

Intended populations/conditions. Patients undergoing physical therapy for various knee pathologies, such as ligament/meniscal injury, osteoarthritis (OA), and patellofemoral pain (43–45). It is applicable for patients undergoing a variety of orthopedic knee procedures and young athletic subjects as well as older adults (46,47).

Version. Although originally described as a single index with 17 items (43), shorter versions have been widely used. A version using Likert-type scales is also available (48).

Content. Single index with 2 sections pertaining to symptoms (pain, crepitus, stiffness, swelling, instability/slipping, buckling, and weakness) and functional limitations (difficulty walking on level surfaces, use of walking aids, limping, going up and down stairs, standing, kneeling, squatting, sitting, and rising from a sitting position) (43,48). A separate scale has been developed to assess sporting activities (43).

Number of items. The original version comprised 17 items (7 for symptoms, 10 for function), but a 14-item version (6 for symptoms, 8 for function) is also used (43,48).

Response options/scale. Patients rate items using descriptive responses, which are translated to a numerical ordinal scale for scoring. Responses for each item are scored from 0–5, with the exception of item 9 (0–3) and item 10 (0–2) in the 17-item questionnaire.

Recall period for items. 1–2 days.

Endorsements. None.

Examples of use. Conditions: anterior cruciate ligament (ACL) injury, cartilage lesions, patellofemoral pain syndrome (PFPS), knee dislocation, and OA. Interventions:

physical therapy, knee braces, ACL reconstruction, autologous chondrocyte implantation, patellar realignment surgery, and total knee replacement (TKR).

Practical Application

How to obtain. Presented in full as an appendix in the original publication (43).

Method of administration. Patient-completed questionnaire. It has not been validated for interview administration (in person or via telephone).

Scoring. The total score is calculated as the sum of scores from the responses to each item, and then transformed to a percentage score by dividing by the maximum total possible score and multiplying by 100 (43,48).

Missing values. While there are no instructions provided as to handling missing data, the original publication only analyzed questionnaires with no missing data (43).

Score interpretation. Possible transformed score range 0–100, where 100 = no knee-related symptoms or functional limitations.

Normative values. Not available.

Respondent burden. It takes approximately 5 minutes to complete the KOS-ADL questionnaire (43). No training or assistance is required as the KOS-ADL is self-explanatory.

Administrative burden. The total score can be calculated in <5 minutes. No training is required for interpretation.

Translations/adaptations. The KOS-ADL instrument has been validated after translation to German (49), Portuguese (50), Turkish (51), and Greek (52).

Psychometric Information

Method of development. Initial item selection was conducted by review of existing patient-reported outcomes (e.g., Cincinnati Knee Scale, Lysholm Knee Scoring Scale, and Western Ontario and McMaster Universities Osteoarthritis Index [WOMAC]) and International Knee Documentation Committee guidelines. The list of items was modified by 12 physical therapists specialized in rehabilitation of musculoskeletal diseases of the knee (43).

Acceptability. No floor effects have been detected (46,47). Acceptable ceiling effects have been reported in people with a variety of knee pathologies undergoing physical therapy and orthopedic surgeon evaluation (43,47). However, high ceiling effects have been reported 6 months after TKR (46).

Reliability. In patients with mixed knee pathologies, the KOS-ADL has demonstrated adequate internal consistency across multiple languages, as well as adequate test–retest reliability for use in groups and individuals (Table 1).

Validity. *Face and content validity.* During development, the KOS-ADL was examined by orthopedic surgeons and physical therapists, who thought that it adequately covered the range of functions/painful activities performed in daily life, ensuring face validity (43). However, since item selection did not involve patient input, this instrument may lack content validity if the instruments from which items were drawn were not themselves derived from patient input (43).

Construct validity. The KOS-ADL shows good correlation with other knee-specific scales, such as the Lysholm Knee Scoring Scale (43), WOMAC subscales (46), and global assessment of function (43). Higher correlations with the physical than mental component score of the Short Form 12 indicates convergent and divergent construct validity (46).

Ability to detect change. The KOS-ADL demonstrates an ability to detect change in patients with a variety of knee disorders (Table 2). Among patients undergoing physical therapy for various knee pathologies, small effect sizes were reported at 1 week, and large effect sizes were reported at 4 and 8 weeks (43). Moderate effect sizes were reported among patients with PFPS, with a minimum clinically important difference of 7.1 (45). Large effect sizes have been reported following TKR (46). The patient-acceptable symptom state has not been reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The KOS-ADL scale is a reliable and valid instrument that is responsive to change in patients with a variety of knee conditions who are undergoing physical therapy or orthopedic procedures.

Caveats and cautions. The lack of direct patient input into item selection means that content validity cannot be assumed. The KOS-ADL uses more descriptive responses to each item as compared to other patient-reported outcomes, which may be confusing or overwhelming for some patients, particularly those with reading difficulties. By design, the KOS-ADL does not include items pertaining to athletic activities, such as running and jumping.

Clinical usability. The KOS-ADL is sufficiently reliable to allow use in individuals with a variety of knee disorders.

Research usability. The KOS-ADL is reliable, valid, and appropriate for measuring change following nonsurgical and surgical interventions in a variety of knee conditions. However, researchers should be aware that if subjects being evaluated are highly physically active, this instrument is not necessarily valid. Researchers should also be consistent with which version of the scale they are utilizing.

LYSHOLM KNEE SCORING SCALE

Description

Purpose. To evaluate outcomes of knee ligament surgery, particularly symptoms of instability (53).

Intended populations/conditions. Patients with knee ligament injury and anteromedial, anterolateral, combined anteromedial/anterolateral, posterolateral rotatory, or straight posterior instability (53).

Version. First published in 1982 (53). The revised version (1985) added an item regarding knee locking, removed items regarding pain on giving way, swelling with giving way, and the objective measure of thigh atrophy, and also removed the reference to walking, running, and jumping above the sections regarding instability, pain, and swelling (54).

Content. The original scale included 8 items: 1) limp; 2) support; 3) stair climbing; 4) squatting; 5) walking, running, and jumping; and 6) thigh atrophy (53). The revised scale also includes 8 items: 1) limp, 2) support, 3) locking, 4) instability, 5) pain, 6) swelling, 7) stair climbing, and 8) squatting (54).

Number of items. 8 items.

Response options/scale. Individual items are scored differently, using individual scoring scales. The revised scale modified the original scoring slightly: 1) limp (0, 3, 5), 2) support (0, 2, 5), 3) locking (0, 2, 6, 10, 15), 4) instability (0, 5, 10, 15, 20, 25), 5) pain (0, 5, 10, 15, 20, 25), 6) swelling (0, 2, 6, 10), 7) stair climbing (0, 2, 6, 10), and 8) squatting (0, 2, 4, 5) (54).

Recall period for items. Not specified.

Endorsements. None.

Examples of use. Conditions: knee ligament injury (anterior cruciate ligament [ACL], posterior cruciate ligament [PCL], medial collateral ligament [MCL], lateral collateral ligament [LCL]), meniscal tears, knee cartilage lesions, osteochondritis dissecans, traumatic knee dislocation, patellar instability, patellofemoral pain, and knee osteoarthritis. Interventions: knee arthroscopy, ligament reconstruction (ACL, PCL, MCL, LCL), meniscal repair, meniscectomy, microfracture, osteochondral autografts, high tibial osteotomy, patellar realignment and stabilization surgery, lateral release, intraarticular hyaluronic acid injection, and therapeutic exercise.

Practical Application

How to obtain. The revised version is freely available in the publication (54). Multiple web sites publish versions of the scale, although they tend to differ slightly.

Method of administration. Original and revised scales were intended for in-person clinician administration (administered by the orthopedic surgeon with the patient's collaboration) (53,54), although subsequent studies have documented using the scale as a patient-completed questionnaire (55). While significantly lower scores have been found for questionnaires versus interview administration, suggesting interview bias (56), 1 study reported a high level of agreement between patients and physiotherapists using a modified version of the Lysholm scale (item for swelling removed) in patients with knee chondral damage (57).

Scoring. Each possible response to each of the 8 items has been assigned an arbitrary score on an increasing scale. The total score is the sum of each response to the 8 items, of a possible score of 100. Computer scoring is not necessary.

Missing values. No instructions provided.

Score interpretation. Possible score range: 0–100, where 100 = no symptoms or disability. Scores are categorized as excellent (95–100), good (84–94), fair (65–83), and poor (≤64) (54).

Normative values. Normative data are available with and without stratification by sex (58,59).

Respondent burden. Time to complete has not been reported, but is expected to vary depending on the administration method (i.e., patient completed versus clinician administered). The Lysholm scale generally uses simple language in its questioning. However, it does use some specific medical terms such as locking, catching, and weight bearing. Administration of this scale as it was intended (i.e., clinician administered) would ensure adequate explanation of such terms, although this may vary between clinicians. As the items relate to everyday tasks, it is not considered that they would have an emotional impact on the individual.

Administrative burden. Less than 5 minutes to score. Training is not necessary, as the scale provides the corresponding score next to each possible response for each item.

Translations/adaptations. Published in English. Although it has been used in international studies, no cross-cultural adaptations have been published.

Psychometric Information

Method of development. Items pertaining to limp, support, stairs, squatting, and thigh atrophy were selected, and items for pain and swelling were adapted from the modified Larson scoring scale (60). The authors added the item for instability, as they deemed this to be an important component of the disability associated with ACL injury (53). The revised scale does not report how the item for locking was selected (54). Four groups of patients were used to compare the original scale to the modified Larson scoring scale: 1) knee ligament injury and anteromedial, anterolateral, and combined anteromedial/anterolateral instability; 2) knee ligament injury and posterolateral rotatory or straight posterior instability; 3) meniscus tears; and 4) chondromalacia patellae (53). Item-response theory was not used in the development of the Lysholm scale.

Acceptability. Rates of missing data have not been reported. There are consistent reports of no floor or ceiling effects (i.e., <15% of patients score the lowest or highest score, respectively) (47,55,61–64).

Reliability. The Lysholm scale appears to have inadequate internal consistency in patients with a variety of knee conditions (Table 1). Test–retest reliability is adequate for use in groups with knee injuries, but is less than adequate for groups with mixed knee pathologies. Reliability may be inadequate for use in individuals. The minimal detectable change has been reported as between 8.9 and 10.1 for knee injuries, while the standard error of the measure is reported to range from 3.2 to 3.6 for knee injuries and from 9.7 to 12.5 for mixed knee pathologies.

Validity. *Face and content validity.* The Lysholm scale has been reported as having face validity, as evaluated by 5 orthopedic surgeons with sports medicine experience (47). Because the items in the Lysholm scale are surgeon derived, content validity from the patient's perspective cannot be assumed.

Construct validity. Multiple studies have reported convergent construct validity for the Lysholm score, finding significant correlations with the Hospital for Special Surgery modified knee ligament rating system, Cincinnati

Knee Ligament Score, International Knee Documentation Committee Subjective Knee Evaluation Form, Fulkerson and Kujala scores, and Western Ontario and McMaster Universities Osteoarthritis Index (63–65). Two studies have reported evidence of convergent and divergent construct validity, finding the Lysholm score to correlate more highly with the Short Form 12 and Short Form 36 physical components than mental components (47,55). The Lysholm score was shown to satisfy the Rasch model after removal of the item for swelling in patients awaiting surgery for knee chondral damage (57).

Ability to detect change. Large effect sizes have been reported following ACL reconstruction (6–9 months postoperative), meniscal repair (1 year postoperative), and microfracture (1–6 years postoperative) (Table 2). Large effect sizes are also reported following 1 month of physical therapy in a group of patients with mixed knee pathologies. The minimum clinically important difference (MCID) and patient-acceptable symptom state (PASS) have not been calculated in any patient population.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The Lysholm scale is a freely available measure that is able to detect change following nonsurgical and surgical intervention. It is considered to have face validity by orthopedic surgeons.

Caveats and cautions. Content validity cannot be assumed, as the items included in the Lysholm scale were surgeon derived. The Lysholm scale was developed as a clinician-administered tool, which increases the potential for interviewer bias if the patient-reported outcome is applied as intended. Despite this, there are inconsistencies between methods of administration of the Lysholm scale in published studies. The MCID and PASS are lacking in psychometric analysis.

Clinical usability. Minimal administrative and respondent burden makes the Lysholm scale attractive for clinical use. The lack of floor and ceiling effects across different knee conditions suggests that the Lysholm scale is useful for tracking improvement with intervention as well as deterioration over time in patients with various knee pathologies. However, clinicians should consider the impact of inadequate reliability in evaluation of individuals.

Research usability. The Lysholm scale is reliable for use in research on ligament and meniscal injuries, chondral injuries, and patellar dislocation. It is important that researchers consistently utilize the same scale version (54). Researchers should be aware that the psychometric properties may change between different administration methods, ensure consistent administration within and between studies, and be aware that clinician and patient ratings may differ substantially. Lack of known MCID is a weakness.

OXFORD KNEE SCORE (OKS)

Description

Purpose. Brief questionnaire for patients undergoing total knee replacement (TKR) that reflected the patient's assessment of their knee-related health status and benefits of treatment (66).

Intended populations/conditions. Patients undergoing TKR.

Version. A new version was proposed on the basis that some surgeons believed that the scoring of the original version was nonintuitive (i.e., lower scores represented better outcome, higher scores represented worse outcome), where the original 12 items are used but the scoring is different (67).

Content. Single index pertaining to knee pain and function (pain severity, mobility, limping, stairs, standing after sitting, kneeling, giving way, sleep, personal hygiene, housework, shopping, and transport).

Number of items. 12 items.

Response options/scale. Each item is followed by 5 responses (scores ranging from 1–5), where 1 = best and 5 = worst outcomes. The modified version also has 5 responses to each item, but the scoring is from 0–4, where 0 = worst and 4 = best outcome.

Recall period for items. Previous 4 weeks.

Endorsements. None.

Examples of use. Conditions: cartilage defects, tibiofemoral osteoarthritis (OA), patellofemoral OA, and rheumatoid arthritis. Interventions: autologous chondrocyte implantation, high tibial osteotomy, unicompartmental knee replacement, and TKR.

Practical Application

How to obtain. The original version can be found in its original publication (66). The modified version is freely available online (www.orthopaedicscore.com/scorepages/oxford_knee_score.html) (67).

Method of administration. Patient-completed questionnaire.

Scoring. Originally, each response to each item was assigned a score from 1–5 (where 1 = no problem and 5 = significant disability). The modified version assigns a score from 0–4 (where 4 = no problem and 0 = significant disability). The total score is calculated as the sum of scores from responses to all 12 items.

Missing values. No instructions provided.

Score interpretation. In the original version, the total score ranges from 12–60 (66), while in the modified version the total score ranges from 0–48 (67). Higher scores in the original version reflect poor outcome and lower scores reflect better outcomes. In the modified version, this is reversed.

Normative values. Not available.

Respondent burden. Reported to involve minimal respondent burden (66). It takes approximately 5–10 minutes to complete the questionnaire. No training or assistance is required since the questions are self-explanatory.

Administrative burden. Scoring is simple and quick (66). Calculation of the total score takes 1–5 minutes. No training is necessary.

Translations/adaptations. Translated and validated in many languages, including Chinese (68), German (69), Japanese (70), Swedish (71), and Thai (72).

Psychometric Information

Method of development. Item generation and reduction was conducted by interviewing patients considering TKR (66).

Acceptability. When tested in patients undergoing TKR, no missing data were reported preoperatively, while postoperative rates of missing data remained low (5%) (66). A more recent study reported no missing data before and 6 months after TKR (46). This study also reported no floor or ceiling effects prior to TKR. Six months postoperatively, although there were no floor effects, there were ceiling effects reported (27% of patients scored the top score).

Reliability. The OKS has adequate internal consistency across multiple languages (66,68–72) (Table 1). The original study reported adequate test–retest reliability for use in groups and individuals (66).

Validity. *Face and content validity*. Extensive input from patients in the development of the OKS ensures content validity.

Construct validity. The OKS shows good correlation with knee-specific and general health questionnaires, such as the Western Ontario and McMaster Universities Osteoarthritis Index, American Knee Society Score, Knee Outcome Survey Activities of Daily Living Scale, and pain and physical function components of the Short Form 36 and Health Assessment Questionnaire (66). Convergent and divergent construct validity is demonstrated by higher correlations with the Short Form 12 physical than mental component (46). The OKS has been shown to fit Rasch models following rescoring of some items (73), and removal of items for limp and kneeling (74).

Ability to detect change. The OKS demonstrates good sensitivity and responsiveness to change (Table 2). Large effect sizes have been reported 6–12 months after TKR (66,75). The OKS has also been found to be a good predictor of revision TKR within 6 months (76). The minimum clinically important difference (MCID) and patient-acceptable symptom state have not been reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The OKS is a self-administered questionnaire developed to measure outcome following TKR. Due to simplicity and ease of administering, it has been used widely, especially in the UK, and is available in languages other than English. For the same reasons, it can be used as a cost-effective screening tool in short-term (<2 years) followup of TKR compared to physician administered instruments, such as the American Knee Society Score, as reported by 1 study (77).

Caveats and cautions. Although simple, some items are "double barreled" and may be confusing to patients (e.g., trouble getting in and out of a car or using public transportation). Some response options potentially overlap with others, which may also cause confusion. The use of an aggregate score combining pain and function may mask changes in 1 domain, particularly given that only 1 of the 12 items relates solely to pain.

Clinical usability. Psychometric testing suggests that the OKS is sufficiently reliable for use in individuals with knee OA. The ease of administration and scoring makes it a useful tool for clinical use. However, clinicians should be aware that some patients may require explanation of individual items, which could introduce interviewer bias.

Research usability. The OKS is a knee OA–specific measure that is reliable, valid, and responsive to change following TKR. Researchers should be aware of the different scoring methods when interpreting findings of previous research. The lack of MCID is a weakness.

WESTERN ONTARIO AND MCMASTER UNIVERSITIES OSTEOARTHRITIS INDEX (WOMAC)

Description

Purpose. To assess the course of disease or response to treatment in patients with knee or hip osteoarthritis (OA) (78,79).

Intended populations/conditions. Patients with knee and hip OA (78,79).

Version. Initially developed in 1982, the WOMAC has undergone multiple revisions (most recent version 3.1). It is available in 5-point Likert, 100-mm visual analog scale (VAS), and 11-box numerical rating scales (80,81). Reduced versions of the WOMAC have been validated but are not endorsed on the WOMAC web site (82–84).

Content. Three subscales: 1) pain severity during various positions or movements, 2) severity of joint stiffness, and 3) difficulty performing daily functional activities.

Number of items. 24 items.

Response options/scale. In the Likert version, each item offers 5 responses: "none" scored as 0, "mild" as 1, "moderate" as 2, "severe" as 3, and "extreme" as 4. Alternatively, the VAS and numerical rating scale versions permit responses to be selected on a 100-mm or 11-box horizontal scale, respectively, with the left end marked as "none" and the right end marked as "extreme" (78,79).

Recall period for items. 48 hours.

Endorsements. Osteoarthritis Research Society International.

Examples of use. Conditions: knee OA, chondral defects, and anterior cruciate ligament (ACL) deficiency. Interventions: physical therapy, massage, self-management, group education, weight loss, exercise, hydrotherapy, Tai Chi, yoga, diet, knee braces, foot orthoses, electrotherapy (e.g., transcutaneous electrical nerve stimulation, laser, pulsed electrical stimulation), acupuncture, pharmacotherapy (drugs, supplements), corticosteroid injection, intraarticular hyaluronic acid injection, arthroscopy, autologous chondrocyte implantation, ACL reconstruction, and total knee replacement (TKR).

Practical Application

How to obtain. Available from Professor Nicholas Bellamy (Australia, e-mail: n.bellamy@uq.edu.au). To obtain licensing and fee information and permission to use the

WOMAC for clinical or research purposes a request needs to be submitted to http://www.womac.org.

Method of administration. Self-administered or interview-administered questionnaire. It has been validated for use in person, over the telephone, or electronically via a computer or mobile phone (79,85–88).

Scoring. The total score for each subscale is the sum of scores for each response to each item, and can be calculated manually or using a computer. The range for possible subscale scores in the Likert format are: pain (0–20; 5 items each scored 0–4), stiffness (2 items, 0–8), and physical function (17 items, 0–68). In the VAS format, the ranges for the 3 subscale scores are: pain, 0–500; stiffness, 0–200; and physical function, 0–1,700 (78,79).

Missing values. If 2 or more pain items, both stiffness items, and 4 or more physical function items are missing, the response should be regarded as invalid and the deficient subscale(s) should not be used in analysis (78).

Score interpretation. Higher scores indicate worse pain, stiffness, or physical function.

Normative values. Australian population-based normative data have been reported, stratified by age and sex (89).

Respondent burden. 5–10 minutes to complete.

Administrative burden. Approximately 5 minutes to score. Training is not necessary.

Translations/adaptations. WOMAC version 3.1 is available in >80 languages (80), and has validated language translations for Arabic (90), Chinese (91), Finnish (92), German (93), Hebrew (94), Italian (95), Japanese (96), Korean (97), Moroccan (98), Singapore (99), Spanish (100), Swedish (101,102), Thai (103), and Turkish (104,105).

Psychometric Information

Method of development. Items were generated by survey of patients with knee or hip OA, review of existing questionnaires (e.g., Health Assessment Questionnaire, Arthritis Impact Measurement Scales), and input from rheumatologists and epidemiologists with experience in clinical assessment of rheumatic diseases. Patients were also utilized in item reduction (78).

Acceptability. The original study and subsequent studies have reported low rates of missing data (46,78). Reports of floor and ceiling effects have differed between studies (46,91,103,105,106). The stiffness subscale has been reported as having floor and ceiling effects prior to intervention (46,91,105). Ceiling effects have been reported by various studies for all subscales 6 months and 2 years after TKR (46,106).

Reliability. The stiffness and function subscales have consistently demonstrated adequate internal consistency in knee OA (Table 1). Studies have generally reported adequate internal consistency for the pain subscale, although there have been reports slightly lower than adequate. There have been mixed findings regarding adequacy of test–retest reliability in knee OA for all subscales. Test–retest reliability for the stiffness subscale may not be adequate for use in individuals with knee OA. One study that investigated test–retest reliability in patients with chondral defects found that all subscales had adequate reliability for use in groups, but only the function subscale was adequate for individual use. The minimal detectable change and standard error of the measure vary according to condition and subscale.

Validity. *Face and content validity.* Since the WOMAC was developed with extensive input from patients with OA, as well as input from academic rheumatologists and epidemiologists experienced in clinical assessment of rheumatologic diseases, the WOMAC can be considered to have face and content validity.

Construct validity. Multiple studies have shown that the WOMAC subscales demonstrate good construct validity. Moderate to strong correlations with measures of similar constructs (e.g., Short Form 36 [SF-36] physical subscales, pain/handicap VAS) suggest convergent construct validity (91,94,95,98,104,105,107,108), while lower correlations with measures such as the SF-36 mental subscales indicate divergent construct validity (91,95,104,105,109). Although Rasch analyses have largely utilized mixed knee and hip OA cohorts, it has been reported that there is no differential item functioning based on affected joint (110). While 1 study found the pain subscale to demonstrate good item separation and unidimensionality in patients with knee or hip OA (111), a subsequent study found that a reduced pain subscale (night pain and pain on standing removed) fit the Rasch model and provided more stable results over time and between patients with knee or hip OA and those who have undergone joint replacement (110). The function subscale demonstrates more variability. Although found to have good item separation and unidimensionality in knee/hip OA, function items for performing light chores, getting in/out of a car, and rising from bed were found to be redundant (111). Similarly, Davis et al (110) suggested a 14-item function subscale, with items for heavy domestic duties, getting in/out of the bath, and getting on/off the toilet removed.

Ability to detect change. The WOMAC appears to be responsive to change following surgical and nonsurgical interventions for knee OA and chondral defects (Table 2). In patients with knee OA, large effect sizes are consistently reported on all 3 subscales up to 2 years post-TKR. Following exercise intervention, the stiffness subscale shows small effect sizes at 2 weeks compared to moderate to large effect sizes for the pain and function subscales; however, these also are small at 6 months. Acupuncture has shown small to moderate effect sizes in the short term (3 weeks), but large effect sizes after 8 weeks. Drug intervention tends to show different patterns across 12 weeks for the 3 subscales. Effect sizes for pain tend to be large initially (1 week), and become more variable at 6 weeks (moderate to large) and 3 months (small to large). In comparison, the stiffness subscale tends to show small to moderate effect sizes over the initial 4 weeks, becoming moderate to large by 3 months. Similarly, effect sizes for function also gradually increase, starting at moderate at 2 weeks, and becoming moderate to large at 6 and 12 weeks. Following surgery for chondral defects, large effect sizes are seen for pain and function 6 and 12 months postoperatively, while moderate effect sizes are seen on the stiffness subscale. The minimum clinically important difference has been calculated for TKR (up to 2 years postoperatively; range for pain

22.9–36, range for symptoms 14.4–21.4, range for function 19–33) and nonsteroidal antiinflammatory use (4 weeks; function 9.1). The patient-acceptable symptom state has been determined to be 31.0 (95% confidence interval 29.4–32.9) for the function subscale in people with knee OA (112).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The WOMAC is one of the most commonly used patient-reported outcomes for knee OA. It is simple and quick to administer and score using guidelines provided. The utilization of patients in development ensures content validity. In addition, the WOMAC has undergone validated translations into multiple languages. The use of individual scores for each subscale, rather than an aggregate score, enhances interpretation.

Caveats and cautions. The need to obtain permission and pay licensing fees prior to use may encourage researchers and clinicians to seek alternatives. The inclusion of tasks in the function subscale that may not be performed regularly by all patients (e.g., stair climbing, taking a bath) may result in missing data. Content validity is not ensured for more physically active patients since the function scale does not include more difficult functional tasks. Rasch analysis suggests that the function subscale contains redundant items.

Clinical usability. The variability in administration methods makes the WOMAC a good choice for clinical use, particularly when dealing with patients with communication difficulties. Minimal floor effects means that the pain and function subscales are able to monitor deterioration in condition over time, while ceiling effects have only been reported following TKR. However, clinicians should consider that the stiffness subscale may not be sufficiently reliable for use in individuals.

Research usability. Psychometric testing indicates that the WOMAC is sufficiently reliable and valid for use in research. The variety of validated language translations and methods of administration is a major strength for WOMAC use in research. A body of research supports the responsiveness to change of the WOMAC following surgical and nonsurgical interventions. Extensive use of the WOMAC in previous research facilitates comparison of new findings.

ACTIVITY RATING SCALE (ARS)

Descriptive

Purpose. Developed as a short, simple, knee-specific questionnaire to evaluate the activity level of patients with various knee disorders who participate in different sports. Intended to provide data on an athlete's highest activity level within the past year (i.e., at a time when they were most active) (113).

Intended populations/conditions. Various knee conditions, including ligament, meniscus, and chondral injury; patellofemoral pain; osteochondritis dissecans; trabecular fracture; and iliotibial band syndrome (113).

Version. No modifications to the original version.

Content. Single index pertaining to frequency of athletic activities: 1) running, 2) cutting, 3) decelerating, and 4) pivoting.

Number of items. 4 items.

Response options/scale. Each item is followed by 5 responses for the frequency of each functional component within the past year.

Recall period for items. 1 year.

Endorsements. None.

Examples of use. Conditions: anterior cruciate ligament (ACL) injury, cartilage injury, and knee osteoarthritis. Interventions: ACL reconstruction, autologous chondrocyte implantation, microfracture, high tibial osteotomy, and total knee replacement.

Practical Application

How to obtain. The ARS can be found as an appendix in the original publication (113).

Method of administration. Patient-completed questionnaire. It has not been validated for interview administration (telephone, in person).

Scoring. Each item is scored from 0–4, where 0 = "less than 1 time a month," 1 = "one time in a month," 2 = "one time in a week," 3 = "two to three times in a week," and 4 = "four or more times in a week." The total score is the sum of scores from responses to each of the 4 items (113).

Missing values. No specific instructions for handling missing values.

Score interpretation. The total possible score range is 0–16, where 16 = more frequent participation.

Normative values. Not available.

Respondent burden. Approximately 1 minute to complete. Respondent burden was intentionally minimized through the inclusion of only 4 items (113).

Administrative burden. Less than 5 minutes to score. No training is required.

Translations/adaptations. None.

Psychometric Information

Method of development. Items were selected by literature review, expert opinion (orthopedic surgeons who specialized in sports medicine, physical therapists, and athletic trainers), and surveying patients with knee disorders. Item reduction involved 50 patients with a variety of knee disorders who were physically active who rated the importance and difficulty associated with each functional task on the preliminary list. The top 4, as agreed by the panel of clinicians, were retained in the final version (113).

Acceptability. Information on missing data and floor/ceiling effects is not available.

Reliability. One study has evaluated the test–retest reliability of the ARS, finding adequate reliability for use in groups and individuals (113) (Table 1). The internal consistency has not been reported.

Validity. *Face and content validity.* The use of patients with knee disorders in both item selection and reduction ensures content validity. Final item selection also involved the opinion of clinicians to ensure face validity (113).

Construct validity. The ARS has been reported to have moderate to strong correlation with other knee-related scales that measure activity levels, such as the Tegner Activity Score, Cincinnati Knee Ligament Score, and Daniel Score, suggesting good convergent construct validity (113).

Ability to detect change. The responsiveness, minimum clinically important difference, and patient-acceptable symptom state have not been reported (Table 2). Rasch analysis was not performed.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The ARS is a short simple measure that represents minimal administrator or respondent burden. As it assesses 4 common components of various sporting activities, rather than nominating specific sports, it is generalizable across a wide range of elite and recreational athletes. In addition, to the extent that activities such as running, stopping, and changing direction are also needed for nonsport activities, it could be applicable to other situations (e.g., work tasks).

Caveats and cautions. Since its focus is limited to specific activities, this scale is most useful as an adjunct to other scales that assess other domains of knee function (114). Other activities such as swimming and jumping cannot be evaluated by this scale. Furthermore, since the ARS does not focus on current ability, but on baseline activity frequency perhaps prior to injury, the validity of the instrument depends on the subject's accurate recollection of this frequency. The accuracy of such recollection may be influenced by the time since injury and by the current state of activity. Lack of evidence for responsiveness to change/sensitivity is also a limitation. The ARS should be used as an adjunct to other knee instruments assessing symptoms and difficulty (113).

Clinical usability. The ARS is a short activity-specific questionnaire, making it good for clinical use. It would be suitable for patients who participate in land-based sports or activities that do not involve jumping as a primary movement. Clinicians should consider that the 1-year recall period may be difficult for some patients.

Research usability. The lack of psychometric data for the ARS limits its use in research. As the scale measures the highest level of activity over the past year, without taking into account time of injury, it may be more suited for within-subject study designs, rather than comparing ratings between subjects.

TEGNER ACTIVITY SCORE (TAS)

Description

Purpose. To provide a standardized method of grading work and sporting activities (54). Developed to complement the Lysholm scale, based on observations that limitations in function scores (Lysholm) may be masked by a decrease in activity level (54).

Intended populations/conditions. Intended for use in conjunction with the Lysholm Knee Scoring Scale, originally in patients with anterior cruciate ligament (ACL) injury (54).

Version. Although in some circumstances it has been modified slightly to accommodate different populations, the standard TAS remains in its original format.

Content. Graduated list of activities of daily living, recreation, and competitive sports. The patient selects the level of participation that best describes their current level of activity.

Number of items. One item is selected from a list of 11.

Response options/scale. A score of 10 is assigned based on the level of activity that the patient selects. A score of 0 represents "sick leave or disability pension because of knee problems," whereas a score of 10 corresponds to participation in national and international elite competitive sports (54). Activity levels 6–10 can only be achieved if the person participates in recreational or competitive sport.

Recall period for items. Current ability.

Endorsements. None.

Examples of use. Conditions: knee ligament injury (ACL, posterior cruciate ligament [PCL], medial collateral ligament [MCL], lateral collateral ligament [LCL]), meniscal tears, knee cartilage lesions, osteochondritis dissecans, traumatic knee dislocation, patellar instability, patellofemoral pain, and knee osteoarthritis (OA). Interventions: knee arthroscopy, ligament reconstruction (ACL, PCL, MCL, LCL), meniscal repair, meniscectomy, microfracture, osteochondral autografts, high tibial osteotomy, patellar realignment and stabilization surgery, lateral release, intraarticular hyaluronic acid injection, and therapeutic exercise.

Practical Application

How to obtain. Freely available in the original publication (54).

Method of administration. Originally established as an in-person, clinician-administered tool (115), but has been used more recently as a patient-completed questionnaire (55,116).

Scoring. A score of 10 is assigned based on the level of activity that the patient selects as best representing their current activity level. Computer scoring is not necessary.

Missing values. Not applicable (single score).

Score interpretation. Possible score range: 0–10. Higher scores represent participation in higher-level activities.

Normative values. Normative data have been presented by sex and age group (58).

Respondent burden. Reported to take mean ± SD 3.3 ± 0.6 minutes to complete in those who have undergone total knee replacement (117). The scale classifies work, recreational, and sport activities in a graded activity scale, using common terminology. As such, patients should not have difficulty selecting which level corresponds to their current activity. Degree of difficulty (measured on a visual analog scale) has been reported to increase with age ($r = 0.25$, $P = 0.03$) (117).

Administrative burden. Scoring time is negligible, as the score is based on a single selected item. Training is not necessary.

Translations/adaptations. Available in English. Although it has been used in international studies, no cross-cultural adaptations have been published. Use in other rheumatology populations has consisted of ankle and shoulder disorders.

Psychometric Information

Method of development. Orthopedic surgeons selected items they believed to be difficult for patients with ACL injury. Forty-three patients with ACL-deficient knees then completed a questionnaire in which they graded these activities according to how difficult they were. This formed the basis of item selection for the TAS.

Acceptability. Studies consistently report no floor or ceiling effects in those with knee injury or OA (i.e., <15% scored lowest or highest score, respectively) (55,61,64,117).

Reliability. The TAS has adequate test–retest reliability for groups with knee injuries and knee OA, although reliability is less than adequate for use in individuals (Table 1). For knee injuries, the minimal detectable change is 1, while the standard error of the measure ranges from 0.4–0.64.

Validity. *Face and content validity.* At face value, the TAS covers a wide variety of activity levels that may be applicable to patients with ACL and other knee injuries. However, as initial activity selection was conducted by orthopedic surgeons, with patient input afterward regarding the difficulty of these selected activities, content validity cannot necessarily be assumed.

Construct validity. Evidence for convergent and divergent construct validity is provided by studies that found higher correlations with the physical component of the Short Form 12 than the mental component (55,61,117). The TAS has also shown significant correlations with the International Knee Documentation Committee Subjective Knee Evaluation Form, Knee Society Score function score, Western Ontario and McMaster Universities Osteoarthritis Index pain and function subscales, and Oxford Knee Score (55,61,64,117).

Ability to detect change. Following meniscal surgery, moderate effect sizes are seen 12 months postoperatively in those with isolated meniscal lesions, and large effect sizes are seen in those with combined lesions (Table 2). In those who have undergone ACL reconstruction, effect sizes are reported to be moderate at 6 months and large at 9 months, 1 year, and 2 years. The minimum clinically important difference (MCID) and patient-acceptable symptom state have not been determined.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The TAS is a simple freely available measure of activity level that spans work, sporting, and recreational activities. It is one of the few patient-reported outcomes that were developed to consider the influence of activity level on other symptoms, such as pain alleviation when aggravating activities are avoided.

Caveats and cautions. The TAS was originally intended and developed for patients with ACL injury as an adjunct to the Lysholm scale, not as a stand-alone measure. The MCID is missing from psychometric analysis. Studies suggest that TAS data need to be adjusted for age and sex (118).

Clinical usability. Clinicians should note that its reliability may be inadequate for use in individuals.

Research usability. Although valid and reliable for use in groups, use of the TAS in research may need to be applied with caution. Given its intent to measure change within patients, the TAS may be more appropriate for within-subject repeated measures studies rather than between-group comparisons.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Garratt AM, Brealey S, Gillespie WJ. Patient-assessed health instruments for the knee: a structured review. Rheumatology (Oxford) 2004; 43:1414–23.
2. Veenhof C, Bijlsma JW, van den Ende CH, van Dijk GM, Pisters MF, Dekker J. Psychometric evaluation of osteoarthritis questionnaires: a systematic review of the literature. Arthritis Rheum 2006;55:480–92.
3. Paxton EW, Fithian DC. Outcome instruments for patellofemoral arthroplasty. Clin Orthop Relat Res 2005;436:66–70.
4. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford: Oxford University Press; 2008.
5. Roos EM, Engelhart L, Ranstam J, Anderson AF, Irrgang J, Marx RG, et al. ICRS recommendation document: patient-reported outcome instruments for use in patients with articular cartilage defects. Cartilage 2011;2:122–36.
6. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293–307.
7. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
8. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63: 737–45.
9. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1988.
10. Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? Ann Rheum Dis 2007;66 Suppl:iii40–1.
11. Irrgang JJ, Anderson AF, Boland AL, Harner CD, Kurosaka M, Neyret P, et al. Development and validation of the International Knee Documentation Committee subjective knee form. Am J Sports Med 2001; 29:600–13.
12. Hefti F, Muller W, Jakob RP, Staubli HU. Evaluation of knee ligament injuries with the IKDC form. Knee Surg Sports Traumatol Arthrosc 1993;1:226–34.
13. Anderson AF. Rating scales. In: Fu FH, Harner CD, Vince KL, editors. Knee surgery. Baltimore: Williams & Wilkins; 1994. p. 275–96.
14. Anderson AF, Irrgang JJ, Kocher MS, Mann BJ, Harrast JJ. The International Knee Documentation Committee Subjective Knee Evaluation Form: normative data. Am J Sports Med 2006;34:128–35.
15. Padua R, Bondi R, Ceccarelli E, Bondi L, Romanini E, Zanoli G, et al. Italian version of the International Knee Documentation Committee subjective knee form: cross-cultural adaptation and validation. Arthroscopy 2004;20:819–23.
16. Metsavaht L, Leporace G, Riberto M, de Mello Sposito MM, Batista LA. Translation and cross-cultural adaptation of the Brazilian version

of the International Knee Documentation Committee subjective knee form: validity and reproducibility. Am J Sports Med 2010;38:1894–9.

17. Fu SN, Chan YH. Translation and validation of Chinese version of International Knee Documentation Committee subjective knee form. Disabil Rehabil 2011;33:1186–9.
18. Haverkamp D, Sierevelt IN, Breugem SJ, Lohuis K, Blankevoort L, van Dijk CN. Translation and validation of the Dutch version of the International Knee Documentation Committee subjective knee form. Am J Sports Med 2006;34:1680–4.
19. Lertwanich P, Praphruetkit T, Keyurapan E, Lamsam C, Kulthanan T. Validity and reliability of Thai version of the International Knee Documentation Committee subjective knee form. J Med Assoc Thai 2008;91:1218–25.
20. Crawford K, Briggs KK, Rodkey WG, Steadman JR. Reliability, validity, and responsiveness of the IKDC score for meniscus injuries of the knee. Arthroscopy 2007;23:839–44.
21. Higgins LD, Taylor MK, Park D, Ghodadra N, Marchant M, Pietrobon R, et al. Reliability and validity of the International Knee Documentation Committee (IKDC) subjective knee form. Joint Bone Spine 2007; 74:594–9.
22. Agel J, LaPrade RF. Assessment of differences between the modified Cincinnati and International Knee Documentation Committee patient outcome scores: a prospective study. Am J Sports Med 2009;37: 2151–7.
23. Greco NJ, Anderson AF, Mann BJ, Cole BJ, Farr J, Nissen CW, et al. Responsiveness of the International Knee Documentation Committee subjective knee form in comparison to the Western Ontario and McMaster Universities Osteoarthritis Index, modified Cincinnati Knee Rating System, and Short Form 36 in patients with focal articular cartilage defects. Am J Sports Med 2010;38:891–902.
24. Irrgang JJ, Anderson AF, Boland AL, Harner CD, Neyret P, Richmond JC, et al. Responsiveness of the International Knee Documentation Committee subjective knee form. Am J Sports Med 2006;34:1567–73.
25. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS): development of a self-administered outcome measure. J Orthop Sports Phys Ther 1998; 28:88–96.
26. Paradowski PT, Bergman S, Sunden-Lundius A, Lohmander LS, Roos EM. Knee complaints vary with age and gender in the adult population: population-based reference data for the Knee injury and Osteoarthritis Outcome Score (KOOS). BMC Musculoskelet Disord 2006;7:38.
27. Roos EM, Roos HP, Ekdahl C, Lohmander LS. Knee injury and Osteoarthritis Outcome Score (KOOS): validation of a Swedish version. Scand J Med Sci Sports 1998;8:439–48.
28. Roos EM, Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS): validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes 2003;1:17.
29. Xie F, Li SC, Roos EM, Fong KY, Lo NN, Yeo SJ, et al. Cross-cultural adaptation and validation of Singapore English and Chinese versions of the Knee injury and Osteoarthritis Outcome Score (KOOS) in Asians with knee osteoarthritis in Singapore. Osteoarthritis Cartilage 2006;14:1098–103.
30. De Groot IB, Favejee MM, Reijman M, Verhaar JA, Terwee CB. The Dutch version of the Knee Injury and Osteoarthritis Outcome Score: a validation study. Health Qual Life Outcomes 2008;6:16.
31. Ornetti P, Parratte S, Gossec L, Tavernier C, Argenson JN, Roos EM, et al. Cross-cultural adaptation and validation of the French version of the Knee injury and Osteoarthritis Outcome Score (KOOS) in knee osteoarthritis patients. Osteoarthritis Cartilage 2008;16:423–8.
32. Salavati M, Mazaheri M, Negahban H, Sohani SM, Ebrahimian MR, Ebrahimi I, et al. Validation of a Persian-version of Knee injury and Osteoarthritis Outcome Score (KOOS) in Iranians with knee injuries. Osteoarthritis Cartilage 2008;16:1178–82.
33. Goncalves RS, Cabri J, Pinheiro JP, Ferreira PL. Cross-cultural adaptation and validation of the Portuguese version of the Knee injury and Osteoarthritis Outcome Score (KOOS). Osteoarthritis Cartilage 2009; 17:1156–62.
34. Chaipinyo K. Test-retest reliability and construct validity of Thai version of Knee Osteoarthritis Outcome Score (KOOS). Thai J Phys Ther 2009;31:67–76.
35. Paker N, Bugdayci D, Sabirli F, Ozel S, Ersoy S. Knee Injury and Osteoarthritis Outcome Score: reliability and validation of the Turkish version. Turkiye Klinikleri J Med Sci 2007;27:350–6.
36. Bekkers JE, de Windt TS, Raijmakers NJ, Dhert WJ, Saris DB. Validation of the Knee Injury and Osteoarthritis Outcome Score (KOOS) for the treatment of focal cartilage lesions. Osteoarthritis Cartilage 2009; 17:1434–9.
37. Salavati M, Akhbari B, Mohammadi F, Mazaheri M, Khorrami M. Knee injury and Osteoarthritis Outcome Score (KOOS): reliability and validity in competitive athletes after anterior cruciate ligament reconstruction. Osteoarthritis Cartilage 2011;19:406–10.
38. Comins J, Brodersen J, Krogsgaard M, Beyer N. Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re-evaluation. Scand J Med Sci Sports 2008;18:336–45.
39. Perruccio AV, Stefan Lohmander L, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS): an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:542–50.
40. Ornetti P, Perruccio AV, Roos EM, Lohmander LS, Davis AM, Maillefert JF. Psychometric properties of the French translation of the reduced KOOS and HOOS (KOOS-PS and HOOS-PS). Osteoarthritis Cartilage 2009;17:1604–8.
41. Goncalves RS, Cabri J, Pinheiro JP, Ferreira PL, Gil J. Reliability, validity and responsiveness of the Portuguese version of the Knee injury and Osteoarthritis Outcome Score-Physical Function Shortform (KOOS-PS). Osteoarthritis Cartilage 2010;18:372–6.
42. Davis AM, Perruccio AV, Canizares M, Hawker GA, Roos EM, Maillefert JF, et al. Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. Osteoarthritis Cartilage 2009;17:843–7.
43. Irrgang JJ, Snyder-Mackler L, Wainner RS, Fu FH, Harner CD. Development of a patient-reported measure of function of the knee. J Bone Joint Surg Am 1998;80:1132–45.
44. Marx R. Knee rating scales. Arthroscopy 2003;19:1103–8.
45. Piva SR, Gil AB, Moore CG, Fitzgerald GK. Responsiveness of the activities of daily living scale of the knee outcome survey and numeric pain rating scale in patients with patellofemoral pain. J Rehabil Med 2009;41:129–35.
46. Impellizzeri F, Mannion A, Leunig M, Bizzini M, Naal F. Comparison of the reliability, responsiveness, and construct validity of 4 different questionnaires for evaluating outcomes after total knee arthroplasty. J Arthroplasty 2010. E-pub ahead of print.
47. Marx RG, Jones EC, Allen AA, Altchek DW, O'Brien SJ, Rodeo SA, et al. Reliability, validity, and responsiveness of four knee outcome scales for athletic patients. J Bone Joint Surg Am 2001;83-A:1459–69.
48. Irrgang J. Development of a health related quality of life instrument to assess physical function related to pathology and impairment of the knee. Pittsburgh: University of Pittsburgh; 1999.
49. Bizzini M, Gorelick M. Development of a German version of the knee outcome survey for daily activities. Arch Orthop Trauma Surg 2007; 127:781–9.
50. Goncalves R, Cabri J, Pinheiro J. Cross-cultural adaptation and validation of the Portuguese version of the Knee Outcome Survey-Activities of Daily Living Scale (KOS-ADLS). Clin Rheumatol 2008; 27:1445–9.
51. Evcik D, Ay S, Ege A, Turel A, Kavuncu V. Adaptation and validation of Turkish version of the Knee Outcome Survey-Activities for Daily Living Scale. Clin Orthop Relat Res 2009;467:2077–82.
52. Kapreli E, Panelli G, Strimpakos N, Billis E, Zacharopoulos A, Athanasopoulos S. Cross-cultural adaptation of the Greek version of the Knee Outcome Survey-Activities of Daily Living Scale (KOS-ADLS). Knee 2010. E-pub ahead of print.
53. Lysholm J, Gillquist J. Evaluation of knee ligament surgery results with special emphasis on use of a scoring scale. Am J Sports Med 1982;10:150–4.
54. Tegner Y, Lysholm J. Rating systems in the evaluation of knee ligament injuries. Clin Orthop Relat Res 1985;198:43–9.
55. Briggs KK, Lysholm J, Tegner Y, Rodkey WG, Kocher MS, Steadman JR. The reliability, validity, and responsiveness of the Lysholm Score and Tegner Activity Scale for anterior cruciate ligament injuries of the knee: 25 years later. Am J Sports Med 2009;37:890–7.
56. Hoher J, Bach T, Munster A, Bouillon B, Tiling T. Does the mode of data collection change results in a subjective knee score? Self-administration versus interview. Am J Sports Med 1997;25:642–7.
57. Smith HJ, Richardson JB, Tennant A. Modification and validation of the Lysholm Knee Scale to assess articular cartilage damage. Osteoarthritis Cartilage 2009;17:53–8.
58. Briggs KK, Steadman JR, Hay CJ, Hines SL. Lysholm score and Tegner activity level in individuals with normal knees. Am J Sports Med 2009;37:898–901.
59. Demirdjian AM, Petrie SG, Guanche CA, Thomas KA. The outcomes of two knee scoring questionnaires in a normal population. Am J Sports Med 1998;26:46–51.
60. Oretorp N, Gillquist J, Liljedahl SO. Long term results of surgery for non-acute anteromedial rotatory instability of the knee. Acta Orthop Scand 1979;50:329–36.
61. Briggs KK, Kocher MS, Rodkey WG, Steadman JR. Reliability, validity and responsiveness of the Lysholm Knee Score and the Tegner Activity Scale for patients with meniscal injury of the knee. J Bone Joint Surg Am 2006;88:698–705.
62. Heintjes EM, Bierma-Zeinstra SM, Berger MY, Koes BW. Lysholm scale and WOMAC index were responsive in prospective cohort of young general practice patients. J Clin Epidemiol 2008;61:481–8.

63. Kocher MS, Steadman JR, Briggs KK, Sterett WI, Hawkins RJ. Reliability, validity, and responsiveness of the Lysholm knee scale for various chondral disorders of the knee. J Bone Joint Surg Am 2004;86-A:1139–45.
64. Paxton EW, Fithian DC, Stone ML, Silva P. The reliability and validity of knee-specific and general health instruments in assessing acute patellar dislocation outcomes. Am J Sports Med 2003;31:487–92.
65. Sgaglione NA, Del Pizzo W, Fox JM, Friedman MJ. Critical analysis of knee ligament rating systems. Am J Sports Med 1995;23:660–7.
66. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg Br 1998;80:63–9.
67. Murray D, Fitzpatrick R, Rogers K, Pandit H, Beard D, Carr A, et al. The use of the Oxford hip and knee scores. J Bone Joint Surg Br 2007;89:1010–4.
68. Xie F, Li S, Lo N, Yeo S, Yang K, Yeo W, et al. Cross-cultural adaptation and validation of Singapore English and Chinese Versions of the Oxford Knee Score (OKS) in knee osteoarthritis patients undergoing total knee replacement. Osteoarthritis Cartilage 2007;15:1019–24.
69. Naal F, Impellizzeri F, Sieverding M, Loibl M, Von Knoch F, Mannion A, et al. The 12-item Oxford Knee Score: cross-cultural adaptation into German and assessment of its psychometric properties in patients with osteoarthritis of the knee. Osteoarthritis Cartilage 2009;17:49–52.
70. Takeuchi R, Sawaguchi T, Nakamura N, Ishikawa H, Saito T, Goldhahn S. Cross-cultural adaptation and validation of the Oxford 12-item knee score in Japanese. Arch Orthop Trauma Surg 2011;131: 247–54.
71. Dunbar M, Robertsson O, Ryd L, Lidgren L. Translation and validation of the Oxford-12 item knee score for use in Sweden. Acta Orthop Scand 2000;71:268–74.
72. Charoencholvanich K, Pongcharoen B. Oxford knee score and SF-36: translation & reliability for use with total knee arthroscopy patients in Thailand. J Med Assoc Thai 2005;88:1194–202.
73. Conaghan PG, Emerton M, Tennant A. Internal construct validity of the Oxford Knee Scale: evidence from Rasch measurement. Arthritis Rheum 2007;57:1363–7.
74. Ko Y, Lo N, Yeo S, Yang K, Yeo W, Chong H, et al. Rasch analysis of the Oxford Knee Score. Osteoarthritis Cartilage 2009;17:1163–9.
75. Garratt A, Brealey S, Gillespie W. Patient-assessed health instruments for the knee: a structured review. Rheumatology (Oxford) 2004;43: 1414–23.
76. Moonot P, Medalla G, Matthews D, Kalairajah Y, Field R. Correlation between the Oxford Knee and American Knee Society scores at mid-term follow-up. J Knee Surg 2009;22:226–30.
77. Medalla GA, Moonot P, Peel T, Kalairajah Y, Field RE. Cost-benefit comparison of the Oxford Knee Score and the American Knee Society Score in measuring outcome of total knee arthroplasty. J Arthroplasty 2009;24:652–6.
78. Bellamy N. WOMAC Osteoarthritis Index user guide. London (Ontario, Canada): University of Western Ontario; 1995.
79. Bellamy N. WOMAC Osteoarthritis Index user guide. Version V. Brisbane (Australia): CONROD, The University of Queensland; 2002.
80. WOMAC-AUSCAN-osteoarthritis global index. URL: http://www.womac.org.
81. Ornetti P, Dougados M, Paternotte S, Logeart I, Gossec L. Validation of a numerical rating scale to assess functional impairment in hip and knee osteoarthritis: comparison with the WOMAC function scale. Ann Rheum Dis 2011;70:740–6.
82. Baron G, Tubach F, Ravaud P, Logeart I, Dougados M. Validation of a short form of the Western Ontario and McMaster Universities Osteoarthritis Index function subscale in hip and knee osteoarthritis. Arthritis Rheum 2007;57:633–8.
83. Whitehouse SL, Lingard EA, Katz JN, Learmonth ID. Development and testing of a reduced WOMAC function scale. J Bone Joint Surg Br 2003;85:706–11.
84. Yang KG, Raijmakers NJ, Verbout AJ, Dhert WJ, Saris DB. Validation of the short-form WOMAC function scale for the evaluation of osteoarthritis of the knee. J Bone Joint Surg Br 2007;89:50–6.
85. Bellamy N, Campbell J, Stevens J, Pilch L, Stewart C, Mahmood Z. Validation study of a computerized version of the Western Ontario and McMaster Universities VA3.0 Osteoarthritis Index. J Rheumatol 1997;24:2413–5.
86. Bellamy N, Campbell J, Hill J. A comparative study of telephone vs on-site completion of the WOMAC 3.0 Osteoarthritis Index. J Rheumatol 2002;29:783–6.
87. Bellamy N, Wilson C, Hendrikz J, Whitehouse SL, Patel B, Dennison S, et al. Osteoarthritis Index delivered by mobile phone (m-WOMAC) is valid, reliable, and responsive. J Clin Epidemiol 2011;64:182–90.
88. Theiler R, Speilberger J, Bischoff H, Bellamy N, Huber J, Kroesen S. Clinical evaluation of the WOMAC 3.0 OA Index in numeric rating scale format using a computerised touch screen version. Osteoarthritis Cartilage 2002;10:479–81.
89. Bellamy N, Wilson C, Hendrikz J. Population-based normative values for the Western Ontario and McMaster (WOMAC) Osteoarthritis Index and the Australian/Canadian (AUSCAN) hand osteoarthritis index functional subscales. Inflammopharmacology 2010;18:1–8.
90. Guermazi M, Poiraudeau S, Yahia M, Mezganni M, Fermanian J, Habib Elleuch M, et al. Translation, adaptation and validation of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) for an Arab population: the Sfax modified WOMAC. Osteoarthritis Cartilage 2004;12:459–68.
91. Xie F, Li SC, Goeree R, Tarride JE, O'Reilly D, Lo NN, et al. Validation of Chinese Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) in patients scheduled for total knee replacement. Qual Life Res 2008;17:595–601.
92. Soininen JV, Paavolainen PO, Gronblad MA, Kaapa EH. Validation study of a Finnish version of the Western Ontario and McMasters University Osteoarthritis Index. Hip Int 2008;18:108–11.
93. Stucki G, Meier D, Stucki S, Michel BA, Tyndall AG, Dick W, et al. Evaluation of a German version of WOMAC (Western Ontario and McMaster Universities) Arthrosis Index. Z Rheumatol 1996;55:40–9. In German.
94. Wigler I, Neumann L, Yaron M. Validation study of a Hebrew version of WOMAC in patients with osteoarthritis of the knee. Clin Rheumatol 1999;18:402–5.
95. Salaffi F, Leardini G, Canesi B, Mannoni A, Fioravanti A, Caporali R, et al. Reliability and validity of the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index in Italian patients with osteoarthritis of the knee. Osteoarthritis Cartilage 2003;11:551–60.
96. Hashimoto H, Hanyu T, Sledge CB, Lingard EA. Validation of a Japanese patient-derived outcome scale for assessing total knee arthroplasty: comparison with Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). J Orthop Sci 2003;8:288–93.
97. Bae SC, Lee HS, Yun HR, Kim TH, Yoo DH, Kim SY. Cross-cultural adaptation and validation of Korean Western Ontario and McMaster Universities (WOMAC) and Lequesne osteoarthritis indices for clinical research. Osteoarthritis Cartilage 2001;9:746–50.
98. Faik A, Benbouazza K, Amine B, Maaroufi H, Bahiri R, Lazrak N, et al. Translation and validation of Moroccan Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index in knee osteoarthritis. Rheumatol Int 2008;28:677–83.
99. Thumboo J, Chew LH, Soh CH. Validation of the Western Ontario and McMaster University Osteoarthritis Index in Asians with osteoarthritis in Singapore. Osteoarthritis Cartilage 2001;9:440–6.
100. Escobar A, Quintana JM, Bilbao A, Azkarate J, Guenaga JI. Validation of the Spanish version of the WOMAC questionnaire for patients with hip or knee osteoarthritis: Western Ontario and McMaster Universities Osteoarthritis Index. Clin Rheumatol 2002;21:466–71.
101. Roos EM, Klassbo M, Lohmander LS. WOMAC osteoarthritis index: reliability, validity, and responsiveness in patients with arthroscopically assessed osteoarthritis. Western Ontario and McMaster Universities. Scand J Rheumatol 1999;28:210–5.
102. Soderman P, Malchau H. Validity and reliability of Swedish WOMAC Osteoarthritis Index: a self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP). Acta Orthop Scand 2000;71:39–46.
103. Kuptniratsaikul V, Rattanachaiyanont M. Validation of a modified Thai version of the Western Ontario and McMaster (WOMAC) Osteoarthritis Index for knee osteoarthritis. Clin Rheumatol 2007;26: 1641–5.
104. Basaran S, Guzel R, Seydaoglu G, Guler-Uysal F. Validity, reliability, and comparison of the WOMAC osteoarthritis index and Lequesne algofunctional index in Turkish patients with hip or knee osteoarthritis. Clin Rheumatol 2010;29:749–56.
105. Tuzun EH, Eker L, Aytar A, Daskapan A, Bayramoglu M. Acceptability, reliability, validity and responsiveness of the Turkish version of WOMAC osteoarthritis index. Osteoarthritis Cartilage 2005;13:28–33.
106. Escobar A, Quintana JM, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. Osteoarthritis Cartilage 2007;15:273–80.
107. Brazier J, Harper R, Munro J, Walters S, Snaith M. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. Rheumatology (Oxford) 1999;38:870–7.
108. McConnell S, Kolopack P, Davis AM. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. Arthritis Rheum 2001;45:453–61.
109. Bombardier C, Melfi C, Paul J, Hawker G, Wright J, Coyte P. Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. Med Care 1995;33 Suppl: AS131–44.
110. Davis A, Badley E, Beaton D, Kopec J, Wright J, Young N, et al. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. J Clin Epidemiol 2003;56:1076–83.
111. Wolfe F, Kong SX. Rasch analysis of the Western Ontario McMaster

questionnaire (WOMAC) in 2,205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Ann Rheum Dis 1999;58:563–8.

112. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Ann Rheum Dis 2005;64:34–7.

113. Marx R, Stump T, Jones E, Wickiewicz T, Warren R. Development and evaluation of an activity rating scale for disorders of the knee. Am J Sports Med 2001;29:213–8.

114. Rick W. Knee injury outcomes measures. J Am Acad Orthop Surg 2009;17:31–9.

115. Hambly K. The use of the Tegner Activity Scale for articular cartilage repair of the knee: a systematic review. Knee Surg Sports Traumatol Arthrosc 2011;19:604–14.

116. Frobell RB, Roos EM, Roos HP, Ranstam J, Lohmander LS. A randomized trial of treatment for acute anterior cruciate ligament tears. N Engl J Med 2010;363:331–42.

117. Naal FD, Impellizzeri FM, Leunig M. Which is the best activity rating scale for patients undergoing total joint arthroplasty? Clin Orthop Relat Res 2009;467:958–65.

118. Frobell RB, Svensson E, Gothrick M, Roos EM. Self-reported activity level and knee function in amateur football players: the influence of age, gender, history of knee injury and level of competition. Knee Surg Sports Traumatol Arthrosc 2008;16:713–9.

119. Bengtsson J, Mollborg J, Werner S. A study for testing the sensitivity and reliability of the Lysholm knee scoring scale. Knee Surg Sports Traumatol Arthrosc 1996;4:27–31.

120. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. J Clin Epidemiol 2003;56:730–5.

121. Dunbar MJ, Robertsson O, Ryd L, Lidgren L. Appropriate questionnaires for knee arthroplasty: results of a survey of 3,600 patients from The Swedish Knee Arthroplasty Registry. J Bone Joint Surg Br 2001; 83:339–44.

122. Fransen M, Edmonds J. Reliability and validity of the EuroQol in patients with osteoarthritis of the knee. Rheumatology (Oxford) 1999; 38:807–13.

123. Stucki G, Sangha O, Stucki S, Michel BA, Tyndall A, Dick W, et al. Comparison of the WOMAC (Western Ontario and McMaster Universities) Osteoarthritis Index and a self-report format of the self-administered Lequesne-Algofunctional index in patients with knee and hip osteoarthritis. Osteoarthritis Cartilage 1998;6:79–86.

124. Angst F, Ewert T, Lehmann S, Aeschlimann A, Stucki G. The factor subdimensions of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) help to specify hip and knee osteoarthritis: a prospective evaluation and validation study. J Rheumatol 2005;32:1324–30.

125. Chesworth BM, Mahomed NN, Bourne RB, Davis AM. Willingness to go through surgery again validated the WOMAC clinically important difference from THR/TKR surgery. J Clin Epidemiol 2008;61:907–18.

126. Davis AM, Lohmander LS, Wong R, Venkataramanan V, Hawker GA. Evaluating the responsiveness of the ICOAP following hip or knee replacement. Osteoarthritis Cartilage 2010;18:1043–5.

127. Theiler R, Bischoff-Ferrari HA, Good M, Bellamy N. Responsiveness of the electronic touch screen WOMAC 3.1 OA Index in a short term clinical trial with rofecoxib. Osteoarthritis Cartilage 2004;12:912–6.

128. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis 2005;64:29–33.

Summary Table for Knee Function Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
Knee function										
IKDC	Symptoms, sport/daily activities, function; variety of knee conditions	Patient completed	10 min	5 min; manual scoring using guidelines provided	Single score; 0–100 (100 = no symptoms, no limitation with daily/sport activities)	Internal: adequate; test–retest: adequate for groups/individuals with knee injuries	Face: adequate; content: cannot be assumed; construct: adequate	Responsive to change following surgery; MCID for cartilage repair, various knee surgeries	No floor/ceiling effects	No patient input in development; long recall period; missing data; lacking psychometric testing in knee OA; aggregate score may mask deficits in 1 domain; multiple versions available
KOOS	Pain, symptoms, ADL, sport/rec, QOL; posttraumatic knee OA and preceding conditions	Patient completed	10 min	5 min; scoring spreadsheet	5 subscales; 0–100 (100 = no problems)	Internal, test–retest: variable (subscale, condition)	Face: adequate; content: adequate; construct: adequate	Responsive to change across a variety of knee conditions following surgical and nonsurgical interventions; MCID: NR	Substantial psychometric testing and cross-cultural validation; individual rather than aggregate scores	Not validated for interview administration; applicability of sport/rec items in older/less physically active patients
KOOS-PS	Function (ADL, sport/rec); knee OA	Patient completed	2 min	<5 min; conversion table	Single score; 0–100 (100 = no difficulty)	Internal: adequate; test–retest: adequate for groups; less than adequate for individuals	Face: adequate; content: adequate; construct: adequate	Responsive to change following physical therapy and hyaluronic acid injection; MCID: NR	Developed using Rasch analysis; minimal burden	Psychometric testing only in knee OA
KOS-ADL	Symptoms, functional limitations; various knee pathologies (ligament/meniscal injuries, OA, PFP)	Patient completed	5 min	<5 min; manual calculation	Single score; 0–100 (100 = no knee-related symptoms or functional limitations)	Internal: adequate; test–retest: adequate	Face: adequate; content: cannot be assumed; construct: adequate	Responsive to change across a variety of knee disorders and interventions (physical therapy, TKR); MCID for PFP	Reliable and valid	No patient input in development; descriptive responses may be confusing; ensure use of consistent version; may not be appropriate for highly active patients
Lysholm Knee Scoring Scale	Limp, support, locking, instability, pain, swelling, stairs, squatting; knee ligament surgery	In-person clinician administration	Variable depending on administration method	<5 min; manual calculation	Single score; 0–100 (100 = no symptoms or disability)	Internal: inadequate; test–retest: adequate only for groups with knee injuries	Face: adequate; content: cannot be assumed; construct: adequate	Responsive to change following surgery and PT; MCID: NR	Freely available; minimal burden	No patient input in development; risk of interviewer bias; multiple versions available
OKS	Pain, function; patients undergoing TKR	Patient completed	5–10 min	<5 min; manual calculation	Single score; original version 12–60 (lower scores = better outcomes); modified version 0–48 (higher scores = better outcomes)	Internal: adequate; test–retest: adequate	Face: adequate; content: adequate; construct: adequate	Responsive to change following TKR; MCID: NR	Reliable, valid, and responsive for knee OA and TKR; cross-cultural validations	Some "double-barreled" items; use of aggregate score; beware of 2 different scoring methods
WOMAC	Pain, stiffness; function; knee and hip OA	Patient- or interview-administered questionnaire (validated for in-person, telephone, and electronic use)	5–10 min	5 min; manual or computer scoring	3 subscales; range depends on version (Likert, VAS); lower scores indicate less pain, stiffness, and functional deficits	Internal: adequate for stiffness and function, variable for pain; test–retest: variable (subscale, condition)	Face: adequate; content: adequate; construct: adequate	Responsive to change following surgical and nonsurgical interventions for knee OA and chondral defects; MCID for TKR and NSAID use	Variety of validated administration methods; validated translations into multiple languages; individual subscale scores; minimal floor and ceiling effects	Licensing and fees required; applicability of function subscale items; redundant items in pain and function subscales (Rasch analysis)
Activity level										
ARS	Athletic activities; various knee disorders; participation in sport	Patient completed	<5 min	1 min; manual calculation	Single score; 0–16 (16 = more frequent participation)	Internal: NR; test–retest: adequate	Face: adequate; content: adequate; construct: adequate	Responsiveness, MCID: NR	Short and simple; adjunct to other knee function measures; generalizable across a variety of athletic and similar tasks	Recall difficulty; lack of psychometric testing
TAS	Level of sport and work participation; knee ligament injury (with Lysholm)	In-person clinician administration	3.3 min	<1 min; score corresponds to single response selected	Single score; 0–10 (higher scores = participation in higher-level activities)	Internal: N/A; test–retest: adequate (groups), less than adequate (individuals)	Face: adequate; content: cannot be assumed; construct: adequate	Responsive to change following meniscal surgery and ACL reconstruction; MCID: NR	Simple; spans work and sport/rec activities	More suited to measure within-patient change; adjustment for age and sex

* IKDC = International Knee Documentation Committee Subjective Knee Evaluation Form; MCID = minimum clinically important difference; OA = osteoarthritis; KOOS = Knee Injury and Osteoarthritis Outcome Score; ADL = activities of daily living; sport/rec = sport/recreation; QOL = quality of life; NR = not reported; KOOS-PS = Knee Injury and Osteoarthritis Outcome Score Physical Function Scale; KOS-ADL = Knee Outcome Survey Activities of Daily Living Scale; PFP = patellofemoral pain; TKR = total knee replacement; PT = physical therapy; OKS = Oxford Knee Score; WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index; VAS = visual analog scale; NSAID = nonsteroidal antiinflammatory drug; ARS = Activity Rating Scale; TAS = Tegner Activity Score; N/A = not applicable; ACL = anterior cruciate ligament.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S431–S437
DOI 10.1002/acr.20639
© 2011, American College of Rheumatology

Measures of Juvenile Fibromyalgia

Functional Disability Inventory (FDI), Modified Fibromyalgia Impact Questionnaire–Child Version (MFIQ-C), and Pediatric Quality of Life Inventory (PedsQL) 3.0 Rheumatology Module Pain and Hurt Scale

STACY R. FLOWERS[1] AND SUSMITA KASHIKAR-ZUCK[2]

INTRODUCTION

Juvenile fibromyalgia (FM) is a chronic noninflammatory musculoskeletal pain condition typically diagnosed in adolescence. Juvenile FM is characterized by diffuse widespread pain, multiple painful tender points, sleep difficulty, fatigue, and other associated symptoms (1). Juvenile FM is also associated with considerable difficulty in physical, social, and emotional functioning (2–9). At present, there are no specific medical tests or disease markers to diagnose this condition, and assessment of symptoms and their impact is primarily by patient report. As noted by the Outcome Measures in Rheumatology Clinical Trials Fibromyalgia Syndrome Workgroup (10), a multidimensional assessment of FM syndrome is essential. Such an assessment should include measures of pain, fatigue, sleep, overall functioning, and quality of life. In patients with juvenile FM, research studies have traditionally utilized more generic pediatric measures that are applicable for many pain conditions. Only one measure has been specifically modified for use in juvenile FM: the Modified Fibromyalgia Impact Questionnaire–child version. In the following sections, we discuss 3 measures that can be used for assessment in juvenile FM, i.e., the Functional Disability Inventory (FDI), the Modified Fibromyalgia Impact Questionnaire–child version, and the Pediatric Quality of Life 3.0 Rheumatology Module Pain and Hurt scale. Measures used to assess pain characteristics, fatigue, and sleep used in pediatric pain disorders, including juvenile FM, are discussed in detail in the Measures of Pathology and Symptoms section in this issue.

FUNCTIONAL DISABILITY INVENTORY (FDI)

Description

Purpose. To measure functional disability (impairment in physical and psychosocial functioning due to physical health status) in children and adolescents with chronic pain. The FDI was initially developed to assess functional disability in children and adolescents (ages 8–18 years) with chronic abdominal pain (11,12), but has subsequently been used with a wide variety of pediatric pain conditions, including juvenile fibromyalgia (FM). The original self-report scale was developed in 1991 and it has no updates or revisions. A parent-report version of the FDI is also available.

Content. The FDI assesses difficulty completing daily activities in home, school, recreational, and social domains such as "completing chores," "being at school all day," "walking the length of a football field," and "doing something with a friend." Items are rated in terms of difficulty the child or adolescent has completing each activity.

Number of items. There are a total of 15 items on the measure with no subscales.

Response options/scale. The child or adolescent rates the amount of difficulty they have completing each activity on a 5-point Likert scale (where 0 = no trouble, 1 = a little trouble, 2 = some trouble, 3 = a lot of trouble, or 4 = impossible).

Recall period for items. Respondents are asked to report how much difficulty they had with completing a variety of activities "in the last few days."

Endorsements. The Pediatric Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (PedIMMPACT) guidelines (13) recommend the FDI for the assessment of physical functioning outcomes in clinical trials of pediatric chronic pain.

Examples of use. The FDI has been widely used in clinical research, including studies assessing the relationship between functional disability and psychosocial functioning, as well as clinical trials with a variety of chronic pediatric pain conditions, including juvenile FM (2,6,7,14).

[1]Stacy R. Flowers, PsyD: The Children's Medical Center of Dayton, Dayton, Ohio; [2]Susmita Kashikar-Zuck, PhD: Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, Ohio.

Address correspondence to Susmita Kashikar-Zuck, PhD, Division of Behavioral Medicine and Clinical Psychology, MLC 3015, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229. E-mail: Susmita.Kashikar-Zuck@cchmc.org.

Submitted for publication February 1, 2011; accepted in revised form May 10, 2011.

Practical Application

How to obtain. Dr. Lynn Walker, Professor of Pediatrics and Director of the Division of Adolescent Medicine and Behavioral Science in the Monroe Carell Jr. Children's Hospital at Vanderbilt, 719 Thompson Lane, Suite 36300, Nashville, TN 37204. Copies and permissions can be obtained directly from Dr. Walker and there is no charge for the FDI.

Method of administration. The FDI is a child/adolescent self-report instrument. It can also be administered in an interview format, if needed, for younger children. The scale can be completed in person, by mail, or by phone. The instrument was designed so that followup assessments to monitor patient progress can be conducted by phone interview (12).

Scoring. Item scores range from 0–4. The total FDI score is a sum of all of the items and can be easily hand scored. Computer scoring is not necessary.

Score interpretation. Scores range from 0–60, with higher scores indicating greater functional disability. Recently, clinical reference points were developed to identify 3 categories of disability in pediatric chronic pain, i.e., no/minimal disability (0–12), moderate disability (13–29), and severe disability (≥30), and these can be used for children and adolescents with a variety of pain conditions, including widespread chronic pain (3).

Respondent burden. Completion of the measure generally takes <10 minutes, although the measure may take longer for younger children with reading difficulties who require interview format administration. It has a Flesch reading ease of 89.7 and a Flesch-Kincaid grade level of 3.2.

Administrative burden. Time to administer is ~5–10 minutes and time to score is <5 minutes. No special training is necessary to administer or score the measure.

Translations/adaptations. The FDI has been translated to the following languages for use: Spanish, French, German, Swedish, Dutch, Arabic, Bulgarian, Polish, Afrikaans, Estonian, Croatian, Slovenian, Macedonian, Romanian, Hungarian, Greek, Russian, Hebrew, Finnish, and East Indian languages, including Hindi, Tamil, Gujerati, Kannada, Malayalam, Marathi, and Telugu (Walker L: unpublished observations). To date, no revalidation studies addressing cultural differences have been conducted.

Psychometric Information

Method of development. Items were generated by reviewing and adapting items from existing adult measures of physical and psychosocial functioning, i.e., the Sickness Impact Profile (15) and the Duke–UNC Health Profile (16). Once the items were selected, pilot testing was conducted with children and their parents in a pediatric outpatient clinic and several items were removed and other items reworded (12) to arrive at the content of the final scale.

Acceptability. The FDI was developed for children as young as 8 years old. The language difficulty of the measure is adapted to the typical reading level of children and adolescents. Missing data are not common, as items are easily understood. There are no known floor or ceiling effects, as very few individuals actually score either 0 or 60. Clinic-based studies have shown that children and adolescents with chronic pain generally obtain scores in the moderate range of disability (total scores 13–29) (3), whereas community-based studies show that healthy school-age children report overall FDI scores in the range of 3–8 (17,18).

Reliability. *Evidence for internal consistency.* Cronbach's alpha reliability coefficients for the FDI are high ($\alpha = 0.85–0.92$) (11,12). The mean interitem correlation is 0.38 (12), which is consistent with the broad domain of functioning covered by the items.

Evidence for stability. Test–retest correlations are high at 2-week ($r = 0.80$, $P < 0.001$), 6-week ($r = 0.70$, $P < 0.001$), and 6-month followup ($r = 0.63$, $P < 0.001$) (12).

Validity. *Evidence of content validity.* Concurrent validity was examined by calculating correlations on the FDI with school absences, which is a common proxy for child disability (15). There was a significant correlation between reported FDI scores and the number of school absences ($r = 0.52$, $P < 0.001$) in the 3 months prior to the clinic appointment. Discriminant validity was evaluated by examining whether the FDI could discriminate between 3 diagnostic groups (abdominal pain with organic etiology, recurrent abdominal pain, and healthy controls). The FDI was able to discriminate among the 3 groups ($F[2,97] = 26.40$, $P < 0.001$). Post hoc analyses revealed significantly higher FDI scores for adolescents with recurrent abdominal pain and organic abdominal pain in comparison to healthy controls (12).

Evidence of construct validity. Construct validity has been demonstrated by examining the association between the FDI and other measures of child well-being. A study of 15 children with juvenile FM found that measures of depression, anxiety, fatigue, and pain (Pearson's $r = 0.42–0.58$, $P < 0.05$) all had positive significant correlations with FDI scores (7). A more recent study had similar results with significant positive correlations with measures of depression ($r = 0.45$, $P < 0.01$) and pain ($r = 0.41$, $P < 0.01$) (3).

Evidence of criterion validity. Predictive validity was examined in an abdominal pain population by correlating FDI scores and school absences due to illness during the 3 months following the clinic appointment ($r = 0.44$, $P < 0.001$). The initial FDI scores were also highly correlated with medication use ($r = 0.26$, $P < 0.05$) and somatic symptoms ($r = 0.45$, $P < 0.001$) at 3-month followup (12). There are no studies examining criterion validity specifically in juvenile FM.

Ability to detect change. Studies have reported FDI results as a sensitive indicator of clinical improvement in juvenile FM. Two treatment studies examining the efficacy of cognitive–behavioral therapy in juvenile FM found significant decreases on the FDI posttreatment (2,14).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The FDI is a widely used measure to evaluate functional impairment in adolescents diagnosed with chronic pain. The measure is recommended for use in pediatric pain clinical trials by the PedIMMPACT guidelines (13). It has strong psychometric properties and has been used as a primary outcome measure in several clinical trials for pediatric chronic pain disorders, including juvenile FM (2,14,19).

Caveats and cautions. Until recently, there were no published clinical reference points for direct interpretation of scores based on clinical norms, which presented a challenge for research and clinical use (20,21). However, Kashikar-Zuck and colleagues (3) recently developed clinical reference points for no/minimal, moderate, and severe disability to allow for clinical interpretation of FDI scores.

Clinical usability. The FDI is a reliable and valid measure of functional impairment in children/adolescents diagnosed with juvenile FM in the clinic setting. It has been found to be an efficient and user-friendly tool for routinely tracking patient outcomes throughout the course of treatment, and has been successfully integrated into busy outpatient clinic settings (22). The FDI can be useful in the development of concrete treatment goals for disability reduction in collaboration with patients and their parents. The instrument poses minimal administrative/respondent burden and has not been found to limit clinical use.

Research usability. The FDI has been successfully used in clinical research in juvenile FM (2,6,7,14). The measure is easy to administer, requires minimal administrative/respondent burden, and is a sensitive indicator of treatment efficacy in clinical trials.

MODIFIED FIBROMYALGIA IMPACT QUESTIONNAIRE–CHILD VERSION (MFIQ-C)

Description

Purpose. The MFIQ-C was developed as a brief measure to assess the spectrum of juvenile fibromyalgia (FM) symptoms and the impact of juvenile FM on the physical and emotional functioning of children/adolescents with juvenile FM. It has been used with patients ages 10–20 years (5,8,9,23). The measure is based on the original FIQ (24) and the MFIQ for adults (25), and was adapted for use in children and adolescents by making minor wording changes (i.e., substituting "work" with "school") (8). At this time, there is limited information about the MFIQ-C as it has been used in a very small number of pediatric studies. For a more thorough description of the original adult FIQ measures, please see the fibromyalgia article in the Measures of Pathology and Symptoms section in this issue.

Content. The MFIQ-C measures physical functioning ("Were you able to do your chores around the house?"), how well the patient feels generally ("During the past week, how many days did you feel good?"), and participation in daily activities ("How many days last week did you miss usual daily activities because you were not feeling well?"). The measure also assesses participation in school activities, pain, fatigue, sleep quality, stiffness, anxiety, and depression using visual analog scales (VAS) that measure the degree to which symptoms interfere with daily activities.

Number of items. The MFIQ-C contains 19 items and 9 subscales. The subscales are: physical functioning (items 1–10), overall well-being (item 11), daily activities (item 12), school (items 13), pain (item 14), fatigue (item 15), sleep quality (item 16), stiffness (item 17), anxiety (item 18), and depression (item 19).

Response options/scale. For the physical functioning subscale (first 10 items), respondents are asked how often they were able to complete a variety of daily activities on a 4-point Likert scale with the response options of "always," "most times," "occasionally," or "never." The overall well-being and daily activities scales (2 items) ask respondents to rate the number of days in the week they felt good and how many days they missed daily activities, respectively (response option range 0–7 days). For the remaining 7 items, patients rate the difficulty they have with each symptom on a 0–10-cm VAS ranging from none to most severe.

Recall period for items. Respondents are asked to report the impact of their symptoms over the past 1 week.

Endorsements. There are no endorsements for the use of the MFIQ-C at present.

Examples of use. The MFIQ-C has been used to examine how physical functioning in juvenile FM is affected by coping strategies (8) and social context such as parental pain history and family environment (9). A more recent study found that adolescents with juvenile FM scored significantly higher on the MFIQ-C than matched healthy controls and that family factors and emotional functioning were related to physical functioning as measured by the MFIQ-C (26). A study conducted in an inpatient adolescent psychiatric setting found that patients who had been diagnosed with juvenile FM scored significantly higher on the measure than those without juvenile FM (23).

Practical Application

How to obtain. The original FIQ and MFIQ were developed and validated by Robert Bennett, MD, who can be contacted by e-mail at bennetrob1@comcast.net. The child version (MFIQ-C) was modified from the adult measures by Laura Schanberg, MD, Division of Pediatric Rheumatology, Duke University School of Medicine, Durham, NC 27710; e-mail: schan001@mc.duke.edu. There is no copyright on the measure.

Method of administration. The MFIQ-C is a brief patient self-report questionnaire.

Scoring. Scoring can be done by individual scale or by composite score; however, the composite score format is recommended by the authors of the adult FIQ measure. As in the FIQ, each scale on the MFIQ-C is transformed to a 0–10 scale score using a normalization procedure so that all scores are expressed in similar units (see the fibromyalgia article in the Pathology and Symptoms section in this issue for detailed scoring procedures). A composite score can be calculated by adding the scores on each of the 10 scales to arrive at a score between 0 and 100, where 0 = no impairment and 100 = severe impairment. Computer scoring for the MFIQ-C is not available. Directions for missing values for items 1–10 are to prorate the items by dividing the score by the number of items endorsed.

Score interpretation. A higher score on the MFIQ-C indicates greater impact of juvenile FM symptoms in each domain assessed. Final scores for each scale range from 0–10 and total scores can range from 0–100. There are no published norms available for the MFIQ-C. Two studies of patients with juvenile FM recruited from pediatric outpatient rheumatology settings reported similar mean ± SD scores of 42.0 ± 22.0 (8) and 40.11 ± 14.07 (26). In contrast, healthy comparison controls reported a mean ± SD score of 27.27 ± 14.10 (26).

Respondent burden. The MFIQ-C takes ~5–10 minutes to complete; however, the format and language may be more difficult for younger children (ages <10 years) as the measure was originally developed for adults. It has a Flesch reading ease of 73.9 and a Flesch-Kincaid grade level of 5.6.

Administrative burden. Administration typically takes 5–10 minutes and scoring takes ~10–15 minutes. Some training and familiarity with the measure is required due to the somewhat complicated hand scoring.

Translations/adaptations. It is unknown whether the MFIQ-C has been translated into languages other than English or if other cultural adaptations have been made.

Psychometric Information

Method of development. Items for the MFIQ-C were modified from the adult version of the FIQ and the MFIQ, with the only modification being replacing items referring to "work" with "school." Children and adolescents were not involved in the development of the measure.

Acceptability. Because the MFIQ-C was modified from an adult measure, the readability of the MFIQ-C may be more difficult for children and adolescents. It is unknown whether missing data are common among children and adolescents.

Reliability. There are no known publications on the reliability of the MFIQ-C. Some studies have demonstrated good internal consistency and test–retest reliability in the adult version of the measure.

Validity. The MFIQ has been validated in adults and has been used clinically in a modified form with children and adolescents; however, no validation has been done specifically in the child and adolescent population. In a study by Schanberg et al (8), the MFIQ-C was found to be a better measure of daily functioning for juvenile FM than the physical function scale of the Arthritis Impact Measurement Scales 2. Additionally, the MFIQ-C has been shown to significantly distinguish juvenile FM patients from healthy controls (26).

Ability to detect change. The MFIQ-C has not been used in longitudinal studies or clinical trials; therefore, no information about sensitivity to change is available.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The MFIQ-C measures important juvenile FM–specific elements that other more generic measures do not incorporate. It appears to be a good measure of the impact of FM, including the core domains of pain, sleep, and fatigue. It is unknown whether the measure is appropriate for evaluating interventions or tracking patient progress over time.

Caveats and cautions. The primary weakness of the MFIQ-C is that there is minimal information about its psychometric properties for use in pediatric populations. It has been used in a small number of research studies, but not much is known about the clinical utility of the measure. Moreover, the hand-scoring algorithm for the measure is somewhat cumbersome and no computer scoring is available.

Clinical usability. The MFIQ-C is a brief measure that can be used to assess the impact of symptoms specific to juvenile FM. It can be easily administered in a clinical setting. However, the lack of information about clinical norms or reference points for interpretation of scores limits its clinical utility at this time.

Research usability. The original adult version of the measure was found to have adequate psychometric properties, but there is very little research on the MFIQ-C. There is currently limited information about whether the measure is reliable, valid, or sensitive to change over time in children and adolescents with juvenile FM.

PEDIATRIC QUALITY OF LIFE INVENTORY (PEDSQL) 3.0 RHEUMATOLOGY MODULE PAIN AND HURT SCALE

Description

Purpose. The disease-specific Rheumatology Module of the PedsQL was designed to assess health-related quality of life among children and adolescents with rheumatic conditions such as juvenile idiopathic arthritis, systemic lupus erythematosus, and juvenile fibromyalgia (FM), and is discussed in detail elsewhere in the juvenile idiopathic arthritis article in the Health Status and Quality of Life section in this issue. The pain and hurt scale of the PedsQL 3.0 Rheumatology Module is relevant to assessing juvenile FM symptoms of muscle and joint pain, stiffness, and sleep difficulties. The current measure was developed in 2002 (27) and it has not undergone any revisions or updates. There are different versions for children and adolescents, and a parent-proxy version of the PedsQL 3.0 Rheumatology Module is also available.

Content. The pain and hurt subscale assesses pain, stiffness, and disrupted sleep due to pain.

Number of items. The PedsQL 3.0 Rheumatology Module pain and hurt scale is composed of 4 items. Items are rated in terms of how much of a problem each symptom has been for the child or adolescent in the past month.

Response options/scale. A 5-point Likert scale is used to assess how often each of the items has been a problem (where 0 = never, 1 = almost never, 2 = sometimes, 3 = often, or 4 = almost always).

Recall period for items. The instructions ask how much of a problem each item has been in the past month.

Endorsements. There are no endorsements for this measure at the present time.

Examples of use. Other than published studies on the development of the measure, there are no studies of clinical or research use of the instrument.

Practical Application

How to obtain. James W. Varni, PhD, Professor of Architecture and Medicine, Department of Landscape Architecture and Urban Planning, College of Architecture, Texas A&M University, 3137 TAMU, College Station, TX 77843-3137; e-mail: jvarni@archone.tamu.edu. Copies can be ordered by going to the following web site: http://www.pedsql.org.

The cost of the PedsQL varies based on the type of funding. For nonfunded academic research, the PedsQL may be used with permission from the author at no charge. For funded research, the rates vary depending on the sponsor (government, foundation, or industry-sponsored research) that includes a royalty fee to Dr. James Varni and a distribution fee to MAPI Research Trust. Users may purchase an annual license or pay a license fee per study. A full list of fees can be obtained online at http://pedsql.org/PedsQL-CostStructure.doc.

Method of administration. The instrument is a self-report measure for children (ages 8–12 years) and adolescents (ages 13–18 years). The assessment is patient reported, but younger children may need to have a clinician administer it. Administration is typically completed in person but may also be completed by phone (28).

Scoring. Items are reverse scored and transformed to a 0–100 scale (where 0 = 100, 1 = 75, 2 = 50, 3 = 25, and 4 = 0) so that higher scores indicate better functioning (or less problems in an area). Scale scores are computed by summing the items and dividing by the number of items answered to account for missing data. If >50% of items in a scale are missing, that scale should not be computed. It is acceptable to impute values based on mean scores when there are missing data (27).

Score interpretation. Scores can range from 0–100, and higher scores on the pain and hurt scale reflect fewer problems with pain and other symptoms. Normative values are available for juvenile FM patients from 2 studies that included the Rheumatology Module pain and hurt scales, one comparing juvenile FM with other rheumatic diseases and another comparing the pain and hurt scale for juvenile FM patients to patients with cancer, patients with rheumatic diseases, and healthy controls (27,29). In each of the studies, the juvenile FM patients had significantly greater problems than the comparison groups on the pain and hurt scale.

Respondent burden. The measure takes <10 minutes for children and adolescents to complete. It has a Flesch reading ease of 84.6 and a Flesch-Kincaid grade level of 3.5.

Administrative burden. Time to administer the entire Rheumatology Module is <10 minutes. Scoring takes ~10 minutes and is simple with minimal training necessary.

Translations/adaptations. The PedsQL 3.0 Rheumatology Module is available in English, Spanish, French, German, Italian, Russian, and Slovenian. Cultural adaptations have been made for English for the US and Spanish for the US.

Psychometric Information

Method of development. The PedsQL 3.0 Rheumatology Module was developed based on the authors' research and clinical experiences with patients with rheumatic diseases. Development included a review of the literature, item generation, cognitive interviews, and pretesting and subsequent field testing of the measure in the target population (27). Patients were involved in the development of the measure by patient and parent focus groups and individual focus interviews.

Acceptability. Items were generated based on developmental level and understanding of concepts. The language difficulty of the measure is adapted to the typical reading level of children and adolescents. In a study specifically examining just the pain and hurt scale, the percentage of missing data among children and adolescents with juvenile FM was <1% (29).

Reliability. *Internal consistency.* A study of 231 children and adolescents with rheumatic diseases found high internal consistency reliability for each of the scales on the Rheumatology Module, with Cronbach's α ranging from 0.87–0.90 for the pain and hurt scale (27). A more recent study focused only on juvenile FM patients found somewhat lower internal consistency of the scale (Cronbach's $\alpha = 0.68$) (29).

Validity. *Evidence of content validity.* The inclusion of medical experts, patients, and patient families as part of the development of the PedsQL along with field testing of the measure warrants sufficient evidence for content validity.

Evidence of construct validity. The known groups method was used to determine construct validity of the Rheumatology Module. Two studies showed significant group differences between those diagnosed with juvenile FM and those with other rheumatic diseases on the pain and hurt scale, with juvenile FM patients reporting more pain and hurt (27,29).

Ability to detect change. One study on a small sample (n = 34) of children with rheumatic diseases, including juvenile FM, showed changes in mean scores on the pain and hurt scale over time across 3 treatment sessions. Mean scores for the initial session were 51.47 and steadily increased to 86.11, indicating lower symptom severity, by the third treatment session (27).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PedsQL 3.0 Rheumatology Module is a brief and easy to administer self-report questionnaire. The pain and hurt scale consists of 4 items that are potentially a good indicator of symptom severity in juvenile FM patients. Initial evidence suggests that the pain and hurt scale has adequate psychometric properties and is sensitive to change. The measure allows for comparison between children and adolescents diagnosed with juvenile FM and other rheumatic diseases (i.e., juvenile idiopathic arthritis, systemic lupus erythematosus) as well as other chronic illnesses (e.g., cancer) and healthy children and adolescents.

Caveats and cautions. Other than the initial validation study published by the authors (27,29), there is limited information about the clinical or research utility of the measure. The module has no established clinical cutoffs and it has not been used as an outcome measure in clinical trials.

Clinical usability. The measure is brief, easy to administer and score, and developmentally appropriate for respondents. Unfortunately, its clinical utility has not been tested.

Research usability. The psychometric properties of this measure are strong and have been tested in multiple populations (children, adolescents, parents) with a variety of rheumatic diseases. The data support research use of this measure, but as of yet, few studies have published research related to the PedsQL Rheumatology Module pain and hurt scale in juvenile FM.

AUTHOR CONTRIBUTIONS

Both authors were involved in drafting the article or revising it critically for important intellectual content, and both authors approved the final version to be published.

REFERENCES

1. Yunus MB, Masi AT. Juvenile primary fibromyalgia syndrome: a clinical study of thirty-three patients and matched normal controls. Arthritis Rheum 1985;28:138–45.
2. Degotardi PJ, Klass ES, Rosenberg BS, Fox DG, Gallelli KA, Gottlieb BS. Development and evaluation of a cognitive-behavioral intervention for juvenile fibromyalgia. J Pediatr Psychol 2006;31:714–23.
3. Kashikar-Zuck S, Flowers SR, Verkamp E, Ting TV, Lynch-Jordan AM, Graham TB, et al. Actigraphy-based physical activity monitoring in adolescents with juvenile primary fibromyalgia syndrome. J Pain 2010; 11:885–93.
4. Kashikar-Zuck S, Lynch AM, Graham TB, Swain NF, Mullen SM, Noll RB. Social functioning and peer relationships of adolescents with juvenile fibromyalgia syndrome. Arthritis Rheum 2007;57:474–80.
5. Kashikar-Zuck S, Parkins IS, Graham TB, Lynch AM, Passo M, Johnston M, et al. Anxiety, mood, and behavioral disorders among pediatric patients with juvenile fibromyalgia syndrome. Clin J Pain 2008;24: 620–6.
6. Kashikar-Zuck S, Vaught MH, Goldschneider KR, Graham TB, Miller JC. Depression, coping and functional disability in juvenile primary fibromyalgia syndrome. J Pain 2002;3:412–9.
7. Reid GJ, Lang BA, McGrath PJ. Primary juvenile fibromyalgia: psychological adjustment, family functioning, coping, and functional disability. Arthritis Rheum 1997;40:752–60.
8. Schanberg LE, Keefe FJ, Lefebvre JC, Kredich DW, Gil KM. Pain coping strategies in children with juvenile primary fibromyalgia syndrome: correlation with pain, physical function, and psychological distress. Arthritis Care Res 1996;9:89–96.
9. Schanberg LE, Keefe FJ, Lefebvre JC, Kredich DW, Gil KM. Social context of pain in children with juvenile primary fibromyalgia syndrome: parental pain history and family environment. Clin J Pain 1998;14:107–15.
10. Mease PJ, Clauw DJ, Arnold LM, Goldenberg DL, Witter J, Williams DA, et al. Fibromyalgia syndrome. J Rheumatol 2005;32:2270–7.
11. Claar RL, Walker LS. Functional assessment of pediatric pain patients: psychometric properties of the Functional Disability Inventory. Pain 2006;121:77–84.
12. Walker LS, Greene JW. The Functional Disability Inventory: measuring a neglected dimension of child health status. J Pediatr Psychol 1991; 16:39–58.
13. McGrath PJ, Walco GA, Turk DC, Dworkin RH, Brown MT, Davidson K, et al. Core outcome domains and measures for pediatric acute and chronic/recurrent pain clinical trials: PedIMMPACT recommendations. J Pain 2008;9:771–83.
14. Kashikar-Zuck S, Swain NF, Jones BA, Graham TB. Efficacy of cognitive-behavioral intervention for juvenile primary fibromyalgia syndrome. J Rheumatol 2005;32:1594–602.
15. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787–805.
16. Parkerson GR Jr, Gehlbach SH, Wagner EH, James SA, Clapp NE, Muhlbaier LH. The Duke-UNC Health Profile: an adult health status instrument for primary care. Med Care 1981;19:806–28.
17. Huguet A, Miro J. The severity of chronic pediatric pain: an epidemiological study. J Pain 2008;9:226–36.
18. Vervoort T, Goubert L, Eccleston C, Bijttebier P, Crombez G. Catastrophic thinking about pain is independently associated with pain severity, disability, and somatic complaints in school children and children with chronic pain. J Pediatr Psychol 2006;31:674–83.
19. Eccleston C, Malleson PN, Clinch J, Connell H, Sourbut C. Chronic pain in adolescents: evaluation of a programme of interdisciplinary cognitive behaviour therapy. Arch Dis Child 2003;88:881–5.
20. Palermo TM. Assessment of chronic pain in children: current status and emerging topics. Pain Res Manag 2009;14:21–6.
21. Palermo TM, Long AC, Lewandowski AS, Drotar D, Quittner AL, Walker LS. Evidence-based assessment of health-related quality of life and functional impairment in pediatric psychology. J Pediatr Psychol 2008;33:983–98.
22. Lynch-Jordan AM, Kashikar-Zuck S, Crosby LE, Lopez WL, Smolyansky BH, Parkins IS, et al. Applying quality improvement methods to implement a measurement system for chronic pain-related disability. J Pediatr Psychol 2010;35:32–41.
23. Lommel K, Kapoor S, Bamford J, Melguizo MS, Martin C, Crofford L. Juvenile primary fibromyalgia syndrome in an inpatient adolescent psychiatric population. Int J Adolesc Med Health 2009;21:571–9.
24. Burckhardt CS, Clark SR, Bennett RM. The Fibromyalgia Impact Questionnaire: development and validation. J Rheumatol 1991;18:728–33.
25. Bennett R. The Fibromyalgia Impact Questionnaire (FIQ): a review of its development, current version, operating characteristics and uses. Clin Exp Rheumatol 2005;23 Suppl:S154–62.
26. Kashikar-Zuck S, Lynch AM, Slater S, Graham TB, Swain NF, Noll RB. Family factors, emotional functioning, and functional impairment in juvenile fibromyalgia syndrome. Arthritis Rheum 2008;59:1392–8.
27. Varni JW, Seid M, Smith Knight T, Burwinkle T, Brown J, Szer IS. The PedsQL in pediatric rheumatology: reliability, validity, and responsiveness of the Pediatric Quality of Life Inventory generic core scales and rheumatology module. Arthritis Rheum 2002;46:714–25.
28. Varni JW, Seid M, Kurtin PS. PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 generic core scales in healthy and patient populations. Med Care 2001;39:800–12.
29. Varni JW, Burwinkle TM, Limbers CA, Szer IS. The PedsQL as a patient-reported outcome in children and adolescents with fibromyalgia: an analysis of OMERACT domains. Health Qual Life Outcomes 2007;5:9.

Summary Table for Juvenile Fibromyalgia Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
FDI	To measure functional disability, i.e., difficulty completing daily activities in home, school, recreational, and social domains in pediatric pain conditions, including juvenile FM	Patient self-report; can be administered in interview form or by phone	Minimal burden; <10 minutes to complete. May take longer for younger children (ages 8–10 years)	Time to administer is ~5–10 minutes and <5 minutes to score	Scores range from 0–60, with higher scores indicating greater functional disability	Excellent	Excellent evidence for concurrent, construct, and criterion validity	Excellent evidence from treatment studies showing sensitivity to change	Brief measure with strong psychometric properties Endorsed by PedIMMPACT guidelines for outcomes in clinical trials Strong evidence for use in clinical and research settings	Items are more heavily weighted toward assessing physical function
MFIQ-C	To assess the spectrum of juvenile FM symptoms and impact on physical and emotional functioning	Patient self-report	Minimal burden; <10 minutes to complete Response format and language may be difficult for children ages <10 years	Time to administer is ~5–10 minutes Hand scored and requires knowledge of the scoring algorithm	Scores range from 0–100 with higher scores indicating greater impact of juvenile FM symptoms	None	Some evidence of content and construct validity	–	Only disease-specific measure for juvenile FM Assesses the core domains of pain, sleep, fatigue, and impact of symptoms on functioning Relatively brief measure	Items are a downward extension of the adult FIQ and not specifically developed in a pediatric population Minimal information about psychometric properties Not much is known about clinical utility
PedsQL 3.0 Rheumatology Module pain and hurt scale	Assesses pain, stiffness, and disrupted sleep due to pain	Patient self-report Different versions for children (ages 8–12 years) and adolescents (ages 13–18 years) Can be administered by phone	<10 minutes for the entire Rheumatology Module	Time to administer is <10 minutes Scoring is simple with no special training necessary	Scores range from 0–100. Higher scores reflect fewer problems with pain and other symptoms	Strong evidence for internal consistency reliability	Excellent evidence of content and construct validity	Some evidence for ability to detect change	Brief and easy to administer and score Strong psychometric properties	Limited information about the clinical or research utility of the measure. No published studies other than initial validation studies

* FDI = Functional Disability Inventory; FM = fibromyalgia; PedIMMPACT = Pediatric Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials; MFIQ-C = Modified Fibromyalgia Impact Questionnaire–child version; PedsQL = Pediatric Quality of Life Inventory.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S308–S324
DOI 10.1002/acr.20640
© 2011, American College of Rheumatology

Measures of Disability

Arthritis Impact Measurement Scales 2 (AIMS2), Arthritis Impact Measurement Scales 2-Short Form (AIMS2-SF), The Organization for Economic Cooperation and Development (OECD) Long-Term Disability (LTD) Questionnaire, EQ-5D, World Health Organization Disability Assessment Schedule II (WHODASII), Late-Life Function and Disability Instrument (LLFDI), and Late-Life Function and Disability Instrument-Abbreviated Version (LLFDI-Abbreviated)

MONIQUE A. M. GIGNAC,[1] XINGSHAN CAO,[2] JESSICA MCALPINE,[2] AND ELIZABETH M. BADLEY[1]

INTRODUCTION

Assessing physical disability is of critical importance to arthritis research, care, and policy. In measuring physical disability, researchers acknowledge the need to gauge health in terms of the impact of a condition on a person's ability to perform everyday activities and not just using indices like mortality and the manifestation of disease symptoms. As a result, it is not surprising that a wide range of physical disability measures exists. For example, there are disability scales tailored to specific arthritis diagnoses, including rheumatoid arthritis, osteoarthritis, ankylosing spondylitis, scleroderma, psoriatic arthritis, lupus, and gout; there are region-specific measures of physical disability for knees, hips, the neck, back, and upper extremities, measures that span childhood, adolescence, and adulthood, and domain-specific measures that assess difficulties with diverse aspects of particular roles like paid employment. Many general health status and quality of life measures also include subscales assessing physical disability.

Yet, the term physical disability is one that is often used loosely and interchangeably with a variety of other terms. These include activity limitations, functional limitations, and functional disability. More recently, social functioning has been introduced into the research lexicon. In general, all of these terms reflect the concept of physical disability as one where a physical health condition or disease is evaluated in terms of its impact, difficulties, or limitations on a range of tasks, activities, or roles that are considered typical of everyday life. It is this definition of physical disability that we adopt for our review. However, not every measure taps the full breadth of tasks, activities, and roles encompassed by this definition. For example, early measures of physical disability, labeled activity or functional limitation measures, were primarily aimed at basic aspects of daily living such as eating, bathing, dressing, using the toilet, and household mobility, and were often used to assess independence in older or chronically ill adults. These measures soon were enhanced with items tapping physical disability with instrumental activities of daily life such as shopping, household chores, meal preparation, and community mobility. In measuring instrumental activities, researchers recognized more complex tasks and acknowledged a wide range of personal, social, and environmental factors beyond disease that could influence disability. Moreover, whereas measures of activity limitations were often used in samples of individuals with relatively severe health problems or impairments, measures of instrumental activity limitations were applied to broader cross-sections of the population, including those with less impairment. Most recently, measures of physical disability or disablement have been applied even more broadly to capture a complete range of functional states from body structures and functions to impairments, activities, and participation in society or social functioning in areas such as work, leisure activities, socializing, and intimate relationships. This broad use of the concept of disability is exhibited most clearly in the World Health Organization's (WHO) International Classification of Functioning, Disability, and Health (1).

Because there are numerous measures capturing diverse domains of physical disability, not all are reviewed here, although some are reviewed elsewhere in this issue (e.g., articles on functional limitations, social functioning and

[1]Monique A. M. Gignac, PhD, Elizabeth M. Badley, DPhil: Toronto Western Research Institute, and University of Toronto, Toronto, Ontario, Canada; [2]Xingshan Cao, PhD, Jessica McAlpine, BA: Toronto Western Research Institute, Toronto, Ontario, Canada.

Address correspondence to Monique A. M. Gignac, PhD, Arthritis Community Research & Evaluation Unit, Toronto Western Research Institute, Main Pavilion 10-316, 399 Bathurst Street, Toronto, Ontario, M5T 2S8 Canada. E-mail: gignac@uhnres.utoronto.ca.

Submitted for publication January 31, 2011; accepted in revised form May 9, 2011.

participation, work disability, health status and quality of life, and the large number of versions of the Health Assessment Questionnaire applied to different disease diagnoses, body regions, and ages). In selecting measures for this article, we supplemented the physical disability measures reviewed elsewhere with those that span a cross-section of tasks, activities, and roles that make up daily life. We begin with an early, arthritis-specific, physical disability measure, the Arthritis Impact Measurement Scales 2 (AIMS2) and AIMS2-Short Form. The remaining measures reviewed are not disease specific. However, they allow arthritis researchers to collect data on physical disability, its determinants, and outcomes that are useful for comparing within and across diseases and health conditions. The Organization for Economic Cooperation and Development (OECD) Long-Term Disability (LTD) Questionnaire is one of the first broadly focused measures developed to assess physical disability. It is reviewed here, particularly because it is an early precursor to newer measures like the EQ-5D developed by the European Quality of Life Group and the WHO Disability Assessment Schedule II (WHODASII). We include these latter measures in our review as examples of easy-to-administer questions that are being applied to a wide range of diseases and health states. Finally, we include the Late-Life Function and Disability Instrument (LLFDI), full and abbreviated versions. These measures reflect examples of efforts researchers have made to expand physical disability beyond tasks and activities to include limitations in roles like socializing with others, employment, care of others, leisure, and hobbies.

No one measure reviewed in this article is likely to satisfy the needs of all researchers wanting to measure physical disability. Some measures will be too narrowly focused either in their emphasis on arthritis or in the domains they capture (e.g., basic activities of daily living and not instrumental activities or social roles). Others may provide a broad overview or snapshot of disability, but lack detail that would be sufficient for clinicians in making decisions for patients. For example, some measures use a time frame of today (e.g., EQ-5D) or the last month (e.g., AIMS2, WHODASII); others ask respondents about “typical” difficulties (e.g., OECD LTD Questionnaire, LLFDI). The former time frame can give a relatively accurate picture of disability, but if one’s current disability or disability in the previous month was unusual in some respects, it may not characterize the overall impact of a health condition on a person’s life. The latter time frame results in respondents trying to characterize their disability in terms of what is usual or normal for them. This might be very helpful in getting an overall picture of the impact of a health condition, but it may not be optimal for detecting small changes in health or for use in some kinds of intervention or longitudinal research or when a person’s appraisal of what is normal for them changes or evolves over time. Despite this, the different measures of physical disability capture the impact of arthritis on a broad range of activities and roles that are meaningful to people living with the disease. They are useful in descriptive or surveillance studies identifying areas of need and they have the potential to generate information on the societal impact or burden of disease. All of the measures would benefit from additional testing to examine their predictive validity and responsiveness to change. However, many are promising as outcomes for intervention research.

The interest and importance of measuring a broad range of activities and roles affected by health conditions like arthritis means that we will likely continue to see refinements, greater sophistication, and greater standardization of measures. For example, one innovation of some measures (e.g., WHODASII, EQ-5D) has been the international, collaborative methods used. Numerous countries participated in the design of questions at the outset of the measure’s development. Traditionally, measures have been developed with little or no input from other cultures and then simply translated into other languages. This has sometimes resulted in poorer validity when the measure is applied to diverse samples. Item banks and computerized adaptive testing are also being applied to measures of disability to maximize the information gained from measures while minimizing time and costs of measurement administration. Researchers also are eager to test measures across different diseases for the purposes of comparative studies, to adapt measures to assess the personal and economic cost of disease, and to use physical disability measures for evaluating treatments and interventions. As such, we can anticipate continued improvements in the quality and application of physical disability measures over time.

ARTHRITIS IMPACT MEASUREMENT SCALES 2 (AIMS2)

Description

Purpose. The AIMS2 is an arthritis-specific health status measure that assesses physical functioning, pain, psychological status, social interactions and support, health perceptions, and demographic and treatment information. The AIMS2 has superseded the original AIMS and was revised in 1992 to have greater specificity and sensitivity, and to incorporate client perceptions of performance (2). Development work for the AIMS2 was in patients with rheumatoid arthritis (RA) and osteoarthritis (OA). A short form of the AIMS2 has been developed (3). Information about the AIMS2-Short Form is presented elsewhere in this article.

Content. The physical function component of the AIMS2 (i.e., physical disability) comprises 6 domains: mobility (using transportation, errands, assistance getting around outside the home); walking and bending (vigorous activities, bending, lifting, stooping, climbing stairs); hand and finger function (writing with a pen or pencil, buttoning clothing, opening jars); arm function (putting on a pullover sweater, combing or brushing your hair, reaching); self-care tasks (help with bathing, dressing, using the toilet); and household tasks (meal preparation, housework, laundry). Other domains not described in detail here are symptoms (pain), role (work), social interaction (social activity, family support), and affect (tension, mood).

Number of items. The total AIMS2 has 78 questions. Factor analysis yields a separate physical function component of 28 items. The 28 items capture 6 domains: mobility

(5 items); walking and bending (5 items); hand and finger function (5 items); arm function (5 items); self-care tasks (4 items); and household tasks (4 items).

Response options/scale. Physical function subscales measure trouble (or absence of trouble) with mobility, walking and bending, hand and finger function, and arm function and are assessed on a 5-point Likert-type scale with 1 = all days, 2 = most days, 3 = some days, 4 = few days, and 5 = no days. Subscales measuring self-care and household tasks are assessed with 1 = always, 2 = very often, 3 = sometimes, 4 = almost never, and 5 = never.

Recall period for items. The past month.

Endorsements. There are no known endorsements.

Examples of use. The AIMS2 has been used as an outcome examining the impact of clinical care in RA (4–7), OA (8), psoriatic arthritis (9–12), ankylosing spondylitis (13,14), fibromyalgia (15,16), carpal tunnel syndrome and Colles fracture (17), hemophilia (18–20), and in patients undergoing joint replacement surgery (21).

Practical Application

How to obtain. Available at the following URLs: www.rheumatology.org/practice/clinical/clinicianresearchers/; www.proqolid.org/instruments/arthritis_impact_measurement_scales_aims2. Copyright is held by Boston University, but there is free access. Also available at reference (22).

Method of administration. Self-administered.

Scoring. The AIMS2 User's Guide provides scoring information for the complete scale. Some items are reverse scaled and require recoding prior to scoring. Scores for each of the 6 domains of physical functioning are summed and then converted to a range of 0–10 using a simple mathematical transformation available in the User's Guide. Computer scoring is available. If an item is missing, the average score of the other scale items may be substituted prior to calculation of subscores. Multiple omissions require a case-by-case examination.

Score interpretation. High scores indicate poor health. No cut off values or normative values are available but scale scores may be adjusted to account for comorbidities. AIMS2 scales were originally discussed as 3 or 5 dimensions of health status. However, many studies discuss the measure using 5 dimensions: physical function, symptom, affect, social interaction, and role.

Respondent burden. Approximately 20–25 minutes to complete. It is a lengthy questionnaire, but not burdensome in terms of reading level required or emotional content.

Administrative burden. Scoring by hand completed in approximately 10 minutes; computerized scoring can be completed in seconds. Minimal training required. User's Guide is available online from the Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID) web page provided above.

Translations/adaptations. Available in English, French, Dutch, Swedish, Chinese, Norwegian, Italian, German, Japanese, Spanish, Greek, Hebrew, Portuguese, Turkish, Russian, and Persian. However, authors of some translated versions of the AIMS2 note the need for more psychometric work (23–27). The AIMS2 has been adapted to ankylosing spondylitis (13). Also, an AIMS scale for children and older adults (Geri-AIMS) has been created. However, these latter scales were based on the original AIMS (not the AIMS2) (28,29).

Psychometric Information

Method of development. The AIMS2 was developed to enhance the original AIMS. Original scale items were developed to go beyond disease activity and to measure a broader array of components identified as important to health by the World Health Organization. The original AIMS contained 45 items. In the AIMS2, 35 items were unchanged, 4 were revised, and 6 were deleted. Patients with RA and OA were involved in testing the measure. Subscales were generated using principal components factor analysis. Test–retest reliabilities used intraclass correlation coefficients, Cronbach's alpha, and kappa statistics.

Acceptability. The AIMS2 is easy to complete. Missing data are not noted as a problem. However, floor and ceiling effects have been observed depending on the patient group observed (26,27).

Reliability. Much of the psychometric work available on reliability has used the original AIMS. In Meenan et al (2), within-scale principal component factor analysis found that items in the physical function subscale loaded on a single factor, except for mobility items, which loaded on more than one factor among those with OA. Internal consistency using Cronbach's alpha coefficients ranges from 0.72–0.91 for patients with RA (n = 299) and 0.74–0.96 for patients with OA (n = 109) across the entire 12 scales. Test–retest intraclass correlation coefficients range from 0.78–0.94 over a 2-week period (2). Other studies have found comparable results for internal consistency and test–retest reliability (24,26,30). Examining the 6 components of physical function, Meenan et al (2) report Cronbach's alphas for RA (n = 299) and OA (n = 109), respectively, as mobility level = 0.85 and 0.83, walking and bending = 0.84 and 0.88, hand and finger function = 0.90 and 0.87, arm function = 0.82 and 0.74, self-care tasks = 0.81 and 0.95, and household tasks = 0.88 and 0.81. Intraclass correlation coefficients were calculated on a subset of 45 respondents with RA or OA with a test–retest time frame of 2–3 weeks: mobility = 0.91, walking and bending = 0.92, hand and finger function = 0.94, arm function = 0.92, self-care tasks = 0.81, and household tasks = 0.81.

Validity. The content of the AIMS2 focuses mainly on function and basic tasks of daily living, with less attention given to disability with instrumental activities or social roles. Much of the psychometric work available related to criterion or construct validity has used the original AIMS. The AIMS scales measuring physical functioning were correlated as expected with other measures of function (e.g., Health Assessment Questionnaire [HAQ]) (i.e., criterion validity) and with disease activity (e.g., swollen joint count, pain, erythrocyte sedimentation rate) (i.e., construct validity) (12,31–34). AIMS2 scale scores were significantly associated with areas patients identified as problematic (2); moderate to high correlations ranging from 0.75–0.89 were also found with other measures of disability (e.g., HAQ, Short Form 36 [SF-36]) (2,12,35,36) and low to moderate correlations (0.3–0.5) with measures of disease activity among patients with ankylosing spondylitis and psoriatic arthritis (10,36). The factor structure identified by the scale developers has not been examined as part of validity testing.

Ability to detect change. The AIMS2 was designed to be sensitive to improvements produced by arthritis therapy (2). Physical function scores were found to provide somewhat greater sensitivity to change than the modified HAQ in one study (37) and similar responsiveness in 2 others (4,38). Comparability also exists between the AIMS2 and SF-36 with some studies finding slightly more responsiveness in the SF-36 (38) and others reporting better responsiveness for the AIMS2 (6).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The AIMS2 has been widely used across different types of arthritis diagnoses and exhibits good psychometric properties. It has been used in intervention research as a patient-oriented outcome and demonstrates comparable responsiveness and sensitivity to change as other disability and global health status measures, including the HAQ and SF-36. Use of the physical function component of the AIMS2 along with the other components allows the evaluation of pain and patient perceptions of the broad impact of arthritis on their lives.

Caveats and cautions. The length and time needed to complete the AIMS2 may hinder its use in clinical, community, and population health research. As a disease-specific measure, the AIMS2 is limited in its potential for use in comparative disease studies. In recent years, the AIMS2 has largely been supplanted by other measures of disability, including the AIMS2-Short Form, HAQ, and SF-36. The AIMS2 is somewhat limited in the scope of its questions assessing disability compared to other measures.

Clinical usability. Psychometric evaluation provides some support for the use of the AIMS2 as a clinical outcome in treatment studies. As noted above, administrative burden may limit its clinical use.

Research usability. Psychometric evaluation provides support for the use of the AIMS2 in research with the caveats noted above.

ARTHRITIS IMPACT MEASUREMENT SCALES 2-SHORT FORM (AIMS2-SF)

Description

Purpose. The AIMS2-SF, first published in 1997, is a shortened version of the AIMS2 and is aimed at measuring health status in people with arthritis. The measure asks about physical functioning, pain, psychological status, and social interactions. Items assessing health perceptions, demographics, and treatment information from the AIMS2 were not included. The development work for the AIMS2-SF was in patients with rheumatoid arthritis (RA), although the measure is intended for broad use across different arthritis diagnoses (3).

Content. Items tapping 5 core domains of the AIMS2 were included (i.e., physical functioning, symptoms, social interaction, role, and affect). These 5 core domains are used in some reports of the AIMS2-SF. However, principal components factor analyses excluded the role items and reported on a slightly different 5-factor solution than the core domains. The new domains are also reported in some studies. They are upper-extremity functioning (e.g., buttoning clothing, using a key, writing, reaching, driving), lower-extremity functioning (e.g., walking, vigorous activity, being in bed or a chair most of the day), affect (e.g., tension, nervousness, feeling a burden to others), symptoms (e.g., morning stiffness, pain), and social interaction (e.g., getting together with friends or relatives, enjoying the things you do) (3).

Number of items. There are 26 items, including upper-extremity functioning (8 items, 2 overlap with lower-extremity functioning), lower-extremity functioning (5 items), affect (4 items), symptoms (3 items), social interaction (4 items), and role (2 items).

Response options/scale. 5-point Likert-type scale. Response options depend on the item and are either 1 (all days), 2 (most days), 3 (some days), 4 (few days), and 5 (no days), or 1 (always), 2 (very often), 3 (sometimes), 4 (almost never), and 5 (never).

Recall period for items. The past month.

Endorsements. There are no known endorsements.

Examples of use. The AIMS2-SF has been used as an outcome measure in intervention studies, including exercise and self-management interventions among patients with RA and osteoarthritis (OA) (39–41).

Practical Application

How to obtain. See Guillemin et al (3), Ren et al (42), ten Klooster et al (43), and Haarvardsholm et al (35) for how to obtain free access to AIMS2-SF items and subscale domains.

Method of administration. Self-administered.

Scoring. Scoring is similar to the AIMS2. Some items are reverse scaled and require recoding prior to scoring. Scores for the different domains are summed and can then be converted to a range of 0–10.

Score interpretation. Higher scores indicate poorer health. No cut off or normative values are available.

Respondent burden. Approximately 10 minutes to complete. The questions are not burdensome in terms of the reading level required or their emotional content.

Administrative burden. Scoring by hand takes approximately 10 minutes; computerized scoring can be completed in seconds. Minimal training required.

Translations/adaptations. Languages available include English, French, German, Dutch, Persian, and Russian.

Psychometric Information

Method of development. The number of items in the measure was reduced from the AIMS2 using a Delphi technique with both patients with RA (n = 12) and experts (i.e., rheumatologists, rehabilitation specialists, and methodology experts, n = 19). Patients and experts reached consensus on items critical to the scale concepts and used information from item analysis as a guide. The reduced scale was submitted to principal components analysis to examine the resulting conceptual components and to compare with the AIMS2. Data for psychometric analysis were drawn from a cohort study of 127 patients with RA (3).

Acceptability. The AIMS2-SF is easy and relatively quick to complete. In general, missing data are not reported as a problem with the exception of the role subscale (e.g., in samples with unemployed, disabled, or retired participants). Depending on the joints affected, some floor and ceiling effects have been found, especially in the physical function subscales (i.e., upper- and lower-extremity functioning) (42,44,45). The AIMS2-SF has been identified for potential inclusion as a core set measure for OA (46).

Reliability. Using the AIMS2-SF in samples of RA and OA, internal consistency using Cronbach's alpha coefficients has been good, often ranging from 0.75–0.87. Exceptions have been the social interaction subscale (ranging from 0.32–0.67) and some studies using the role subscale (3,42,44,45,47). Test–retest correlations also have been favorable with intraclass correlations over 2 days to 1 month exceeding 0.80, although lower correlations have been found for the affect and social interaction subscales (3,44,45,47).

Validity. Similar to the AIMS2, the content of the AIMS2-SF, focuses mainly on function and basic tasks of daily living. Little attention is given to disability with instrumental activities or social roles. In general, the AIMS2-SF and AIMS2 had comparable criterion validity with other measures of disability and health status (e.g., modified Health Assessment Questionnaire [MHAQ], Short Form 36 [SF-36], Western Ontario and McMaster Universities OA Index, Disease Activity Score in 28 joints). The physical function subscales of the AIMS2-SF also demonstrate reasonable construct validity and has been found to be significantly associated with greater pain, medication use, lost work days, disease symptoms like joint stiffness, tender joint count, and swollen joint count, and patient and physician global health assessments (3,42,44,45,47). Inconsistent factor structures point to the need for additional testing of subscales in samples with RA and OA (3,42).

Ability to detect change. Additional research is needed using the AIMS2-SF, although preliminary indications suggest no differences between the AIMS2 and AIMS2-SF in responsiveness and comparability to the SF-36 and MHAQ (3,35). Guillemin et al report that the 3-month sensitivity to change was similar to the AIMS2, with the standardized response means at 3 months being high in the physical function and symptom subscales (3). Research by Taal et al also found similar sensitivity to change in the physical function, symptom, and affect components of the AIMS2-SF and the AIMS2 (but less responsiveness in the social interaction and role components). The physical function and symptom components of the AIMS2-SF were more sensitive to change than the MHAQ and visual analog scale (pain) measures (37). The AIMS2-SF has been used as an outcome in an exercise program and self-management intervention (40,41). No significant changes in AIMS2-SF were found. It is not clear whether the measure was not sensitive to change or whether the intervention did not result in meaningful change.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The AIMS2-SF has been shown to have similar psychometric properties as the AIMS2 with the additional benefit of being much shorter. It has recently been identified as a potentially important core measure for assessment of OA disability (46) and has been used in several European intervention studies.

Caveats and cautions. As a disease-specific measure, the AIMS2-SF is limited in its potential for use in comparative disease studies. Different factor loadings and structures may occur when applying the AIMS2-SF to samples of RA and OA patients. Some items may load on more than one factor (e.g., both lower- and upper-extremity functioning). Studies have varied in their use of the role component of the AIMS2-SF. Other measures of work disability and role participation provide more in-depth information on this aspect of disability.

Clinical usability. Psychometric evaluation provides some support for the use of the AIMS2 as a clinical outcome in treatment and intervention studies. However, more research is needed to determine its usefulness as a measure to guide clinical decision making at point of contact with patients.

Research usability. Psychometric evaluation provides support for the use of the AIMS2 in research with the caveats noted above.

THE ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) LONG-TERM DISABILITY (LTD) QUESTIONNAIRE

Description

Purpose. The OECD measure of disability was among the first international efforts to assess the impact of ill health on tasks and activities. It was developed in 1981 to facilitate international comparisons of disability and monitor changes in disability over time across a range of health conditions. Its emphasis is on measuring long-term disruptions to normal activities. Subsequent World Health Organization (WHO) disability questionnaires were based on items from the OECD (48).

Content. Items are combined to assess limitations in activities related to daily living. Various dimensions are covered, including eyesight (e.g., is your eyesight good enough to read ordinary newspaper print), hearing (e.g., can you hear what is said in normal conversation with 3 or 4 other persons), speech (e.g., can you speak without difficulty), upper mobility (e.g., can you carry an object of 5 kilos for 10 meters, can you cut your own food), lower-extremity functioning (e.g., can you walk up and down one flight of stairs without resting), mobility (e.g., walking 400 meters, running 100 meters), and daily activities (e.g., can you dress and undress, can you get in and out of bed).

Number of items. The full measure includes 16 questions. An abbreviated, core set version includes 10 items.

Response options/scale. Responses are on a 4-level scale: yes (without difficulty), yes (with minor difficulty), yes (with major difficulty), and no (not able to do).

Recall period for items. Respondents are asked what they are able to do on a normal day.

Endorsements. There are no known endorsements. The OECD LTD Questionnaire has been supplanted by newer measures (e.g., WHO Disability Assessment Schedule II).

Examples of use. Although the OECD LTD Questionnaire has been largely supplanted by other disability measures, it is worth noting that the scale or some of its items have been used by a number of OECD countries in national population health surveys, including in France, Japan, Germany, the US, and Canada. For example, recent Canadian population health disability surveys used several OECD LTD Questionnaire items. The OECD LTD Questionnaire is recognized as an early precursor to other instruments like the European Quality of Life and the WHO Quality of Life (WHOQOL) questionnaires.

Practical Application

How to obtain. A copy of the questions is available in reference (49).

Method of administration. Questions can be interviewer administered or self-administered.

Scoring. Scoring instructions were unavailable. However, some studies treated each item separately (i.e., did not combine them) or created a summary total of the number of areas with at least some disability by counting items where respondents indicated they had at least some difficulty performing the activity (48).

Score interpretation. No standard scoring availability. Higher levels or counts reflect greater disability.

Respondent burden. Items can be completed in less than 10 minutes.

Administrative burden. The items are simple to administer. However, detailed information on scoring is not available.

Translations/adaptations. English, Dutch, Finnish, French, German, and Japanese. It is not clear whether the OECD LTD Questionnaire has been translated into additional languages.

Psychometric Information

Method of development. Eight countries (Canada, Finland, France, Federal Republic of Germany, The Netherlands, Switzerland, UK, and US) along with the WHO participated in a common development effort aimed at creating a “Healthfulness of Life” core set of questions. The result was the 16-item OECD Questionnaire. It is unclear whether patients were involved. Item response theory was not used in development of the questions.

Acceptability. Floor effects are not uncommon in those under 65 years of age, with many people reporting no difficulty with any of the activities.

Reliability. Test–retest reliability using an 11-item OECD LTD Questionnaire was low with a 2-week interval. Agreement was only between 30–70%. Although a substantial percentage (~50%) of interviews used proxy respondents, further analyses determined that inconsistencies were not due to proxy respondents ([50] as cited in [49]).

Validity. Although the OECD LTD Questionnaire appears to have reasonable face validity, very little systematic validity testing has been carried out on the measure. Canadian data found low to moderate correlations with rehabilitation patients completing the OECD LTD Questionnaire compared to physician mobility ratings, which are only relevant to a small part of the scale items. Correlations ranged from 0.14–0.54 (49). Wijilhuizen and Ooijendijk reported similar findings among Dutch patients (48). Other Canadian studies have looked at the construct validity of selected OECD items in samples of people with arthritis, finding that greater difficulty with OECD items was significantly associated with dependence (i.e., assistance from others) and work disability (51–55).

Ability to detect change. Because the original purpose of the measure was to generate profiles of disability levels in the general population, the OECD LTD Questionnaire has largely not been used to examine change. McDowell cites some sensitivity results for different medical conditions ranging from 61–85%, with the highest sensitivity among those with vision, hearing, and speech problems (49).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The OECD LTD Questionnaire is considered an early attempt to develop internationally applicable disability items. The WHOQOL and EQ-5D are later examples. Although the OECD LTD Questionnaire items are not used as a single measure, individual items continue to be used in some national population studies applicable to arthritis.

Caveats and cautions. The OECD LTD Questionnaire has largely been supplanted by other disability measures.

Clinical usability. The OECD LTD Questionnaire has been supplanted by other disability measures and should not be used in clinical evaluations.

Research usability. Individual items continue to be used in some population health surveys. However, there may be wording variations and the items are not used as a single measure.

EQ-5D

Description

Purpose. The EQ-5D is a generic measure of health status for use in clinical, population, and economic appraisals with adult samples. The measure was developed by the European Quality of Life Group (EuroQol) to act as a core set of items for use in international studies measuring health-related quality of life across a wide range of health conditions and treatments. It provides a simple descriptive profile and single index values of health status. An early measure was first published in 1990 (56) and later finalized as the EQ-5D (57,58).

Content. The EQ-5D has 2 components. A descriptive system (EQ-5D descriptive system) uses single items to assess disability with 5 dimensions: mobility, self-care, usual activities (e.g., work, study, housework, family, or leisure activities), pain/discomfort, and anxiety/depression. In addition, a single visual analog scale (EQ-5D VAS) assesses self-rated health with end points labeled "Best imaginable health" and "Worst imaginable health."

Number of items. 6.

Response options/scale. For each item in the EQ-5D descriptive system, there are 3 levels of response: 1 = no problems (e.g., "I have no problems in walking about," "I have no problems with performing my usual activities," "I have no pain or discomfort"), 2 = some problems (e.g., "I have some problems washing or dressing myself," "I am moderately anxious or depressed"), and 3 = extreme problems (e.g., "I am confined to bed," "I am unable to wash or dress myself," "I am unable to perform my usual activities," "I have extreme pain or discomfort," "I am extremely anxious or depressed") (56–59). EQ-5D VAS scores range from 0 (worst imaginable health state) to 100 (best imaginable health state).

Recall period for items. Today.

Endorsements. There are no known endorsements.

Examples of use. The EQ-5D has been used in a large number of studies across samples with diverse health conditions, as well as population studies (the EuroQol web site reports over 2,200 studies using EQ-5D as of the end of 2010). It has been applied to studies of rheumatoid arthritis (RA) (57,58), osteoarthritis (OA) (60–62), ankylosing spondylitis (63,64), gout (65), juvenile idiopathic arthritis (66), chronic low back pain (67,68), systemic lupus erythematosus (69), and fibromyalgia (70). The EQ-5D has also been used in treatment and intervention studies with psoriatic arthritis (71), RA (72–75), OA (76,77), ankylosing spondylitis (78–80), juvenile idiopathic arthritis (81), total hip/knee replacement surgery (62,82,83), and knee pain (84).

Practical Application

How to obtain. Those using the EQ-5D are asked to register their research online. A copy of the questions and User Guide is available at URL: www.euroqol.org. In terms of cost, whether licensing fees exists is determined by the EuroQol Executive Office and is based on information provided by users (e.g., type of study, sample size, requested languages).

Method of administration. The EQ-5D was designed to be easily self-administered. It can also be interviewer administered face-to-face or by telephone. A proxy version is available as well as a web-based, tablet, and PDA version.

Scoring. Individual profiles are created using the 5 dimensions of the EQ-5D and are called the EQ-5D Health State. For example, a score of 11222 would indicate no difficulties with mobility and self care, but some/moderate problems with usual activities, pain/discomfort, and anxiety/depression. Individual scores can be converted into a summary called the EQ-5D Index. The EQ-5D Index uses a utility-weighted scoring system that has been derived from extensive studies with different countries or by taking into account an individual's own preferences as reflected in the VAS rating scale from 0–100. A constant is also subtracted if one or more dimensions are scored at 2 or 3, and a further constant if one or more dimensions are scored at 3. A negative score is possible in creating the EQ-5D Index, representing a state "worse than death." The EQ-5D was developed using health economics principles. The EuroQol web site provides information on weights derived from EQ-5D VAS scores, as well as weights derived from time trade-off valuation techniques. EQ-5D VAS scores can be converted into quality-adjusted life year (QALY) calculations for economic analyses. Weights have been derived for over 14 countries, with additional weights in development.

Score interpretation. Descriptive patterns for the EQ-5D Health State can result in 243 (i.e., 3^5) possible disability combinations ranging from 11111 to 33333. Weighted scores for the EQ-5D Index indicate 1.0 = the best imaginable health state, 0 = death, as well as negative scores representing a state "worse than death" (22). VAS scores range from 0 (worst imaginable health state) to 100 (best imaginable health state). Researchers have noted that, depending on the item weights and algorithms, widely differing QALY gains and cost-effectiveness estimations may result (84,85–94). Where possible, researchers should use the algorithms specific to their country by consulting the EuroQol web site.

Respondent burden. The EQ-5D was designed to be extremely brief and can be completed in less than 2 minutes. Items are easy to understand and emotional sensitivity of topics is low.

Administrative burden. Time and training are needed to score the EQ-5D. The EuroQol web site must be consulted to register studies and to determine the appropriate rates for a country.

Translations/adaptations. An EQ-5D-Y is available for children (95). More than 120 translations of the EQ-5D exist. Examples of languages include Dutch (and Dutch for Belgium), English (and English for Australia, Canada, New Zealand, UK, US, Singapore, and South Africa), Finnish, French (and French for Belgium, Canada, and Switzerland), German (and German for Austria and Switzerland), Norwegian, Swedish, Spanish (and Spanish for Argentina, Chile, Columbia, Costa Rica, Guatemala, Mexico, Peru, US, Uruguay, and Venezuela), Afrikaans, Indonesian, Bulgarian, Italian, Cantonese for Hong Kong, Japanese, Catalan, Estonian, Latvian, Chinese (and Chinese for Singapore and Taiwan), Lithuanian, Malay, Croatian, Polish, Czech, Portuguese (and Portuguese for Brazil), Danish, Romanian, Russian (and Russian for Israel), Greek, Thai, Hebrew, Slovakian, Turkish, Hungarian, and Slovenian.

Psychometric Information

Method of development. Although originally created by EuroQol beginning in 1987, membership has grown to include members from North America, Asia, Africa, Australia, and New Zealand. The goal of the measure was to create an easy-to-administer core set of items for use in international studies across a wide range of health conditions and treatments. The dimensions were selected after a detailed examination of existing health status measures, including the Quality of Well-Being Scale, Sickness Impact Profile, Nottingham Health Profile, and Rosser Index. The number of health states in each dimension was deliberately kept to a minimum so that the measure could easily be administered and used for decision making.

Acceptability. Reports of missing data are variable. In testing the original EuroQol questionnaire, missing data were reported as much as 40% in the UK sample (56). Another study using the EQ-5D reported less than 1% missing data among returned questionnaires (64). The most frequently omitted items were pain/discomfort and anxiety/depression. Floor effects are not unusual in general populations, with individuals reporting 11111 (i.e., no problems) (22). Among patients with ankylosing spondylitis, floor effects in the 5 dimensions ranged from 10.4% (pain/discomfort) to 61.7% (self-care) and ceiling effects ranged from <1% (mobility) to 20.2% (pain) (64).

Reliability. Test–retest reliability for the EQ-5D Index ranges from intraclass correlation coefficient (ICC) 0.64–0.78 in samples ranging from 1 week to 3 months (58,96). ICC values for the EQ-5D VAS ranged from 0.70–0.85 (58). Gamma coefficients ranged from 0.57–0.80. In a sample of 82 people with knee OA measured twice over a 1-week period, ICC for the EQ-5D Index was 0.70 (95% confidence interval [95% CI] 0.58–0.80) and was 0.73 (95% CI 0.61–0.82) for the EQ-5D VAS (60). Comparing telephone and face-to-face interviews in a sample of older adults, McPhail and colleagues found moderate to high levels of agreement. ICC values were 0.82 for EQ-5D Index scores and 0.58 for the EQ-5D VAS (item kappas ranged 0.67–0.83) (97).

Validity. In studies of patients with RA, the EQ-5D Index was significantly associated with Health Assessment Questionnaire (HAQ), depression, and anxiety. EQ-5D VAS scores were significantly associated with HAQ, self-assessed joint pain, and depression (57,58). Among patients with knee OA, the EQ-5D Index correlated with arthritis duration and greater Western Ontario and McMaster Universities OA Index and Short Form 36 scores, but lacked discriminative ability among those with moderate disability (60). Similarly, in a study of patients with psoriatic arthritis, the EQ-5D Index did not discriminate well among patients with and without disability compared to the Psoriatic Arthritis Quality of Life questionnaire and HAQ (98). Some authors have argued that the EQ-5D lacks dimensions such as dexterity, social functioning, and vitality that are important to disability and that might be responsible for observed gaps in the distribution of EQ-5D Index scores, especially in the midutility range (between 0.30 and 0.5; results in bimodal distributions of scores) (99). This may also be responsible for the ceiling and floor effects noted in a number of studies. Response options and algorithms used to create QALYs have been noted as problematic in a number of studies (70,85–93,100).

Ability to detect change. The EQ-5D has rarely been used as a primary outcome in intervention studies, making it difficult to evaluate its appropriateness to detect change. Cut off points have been suggested for the EQ-5D to identify acceptable health status for RA patients (101). Specifically, the patient acceptable symptom state cut point with 80% specificity was estimated to be 0.70 in the EQ-5D Index. The cut point was 0.65 when the 75th percentile was used. Minimal clinically important improvement cut points assessed by 80% specificity varied from 0.10–0.19 in the EQ-5D Index. In a sample of patients with RA, the EQ-5D Index was found to be significantly associated with changes in HAQ pain, joint pain, depression, and anxiety over 3 months (58). Other studies have found no changes in EQ-5D utility scores in treatment of patients with RA (72). Among patients with ankylosing spondylitis, responsiveness was found to be weak for the EQ-5D Index, but good for the VAS (64).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. It is quick and easy to develop an EQ-5D Health State profile. The EQ-5D has been used in a number of studies with arthritis populations. It allows comparison to other conditions and allows for economic evaluations.

Caveats and cautions. Scoring is complex for the EQ-5D Index. Some authors have argued that there are important dimensions missing and a bimodal distribution of scores may compromise the validity of the measure and its ability to detect change. Researchers have noted that, depending on the item weights and algorithms, widely different QALY gains and cost-effectiveness estimations may result.

Clinical usability. The EQ-5D has been used in clinical settings. It is easy and quick to generate an individual patient profile. However, the measure is not detailed enough to use as a clinical decision making tool.

Research usability. May be useful as core variable to describe populations, but does not provide a lot of detail.

WORLD HEALTH ORGANIZATION DISABILITY ASSESSMENT SCHEDULE II (WHODASII)

Description

Purpose. The WHODASII measures disablement by assessing a range of activity limitations and participation restrictions. It can be used for community health studies, population surveillance, or clinical assessments. The WHODASII is a generic measure that does not target a specific disease but can be used to compare individuals with difficulties stemming from disease, illness, injury, and mental, emotional, or substance abuse problems. The measure was first published in 2000. We used the version published on the WHO web site in 2010 for this review.

Content. Six domains are assessed: cognition (i.e., understanding or communicating, understanding what people say, concentration, remembering, starting a conversation), mobility (i.e., getting around, standing for long periods, moving around the home, getting out of your home, walking a kilometer), self-care (i.e., washing your whole body, getting dressed, eating, staying by yourself for a few days), getting along with people (i.e., maintaining a friendship, getting along with those close to you, making new friends, sexual activities), life activities (i.e., household work, employment, school activities), and participation (i.e., joining in community activities, barriers or hindrances in world around you, living with dignity, taking care of health, emotional impact of health).

Number of items. The 36-item version of the WHODASII includes cognition (6 items), mobility (5 items), self-care (4 items), getting along with people (5 items), life activities: household (4 items), life activities: work or school (4 items), and participation (8 items). Seven additional items in the life activities and participation domains are asked of individuals reporting any difficulties with activities. These items ask about the number of days health problems resulted in missing, reducing time with, or slowing down activities and roles.

Response options/scale. Questions are assessed on a 5-point Likert-type scale, where 1 = none, 2 = mild, 3 = moderate, 4 = severe, and 5 = extreme or cannot do.

Recall period for items. The past 30 days.

Endorsements. Endorsed by the WHO.

Examples of use. The WHODASII has been used in a large number of studies with diverse samples. It has been administered in population and community health surveillance studies, as part of clinical assessments, and in intervention research. In arthritis, the WHODASII has been applied to samples of rheumatoid arthritis (RA), osteoarthritis (OA), scleroderma, and ankylosing spondylitis (102,103).

Practical Application

How to obtain. Available from the WHO web site: http://www.who.int/icidh/whodas/instrument_download.html. There is no cost to the user. Users must register to use the WHODASII.

Method of administration. An interviewer-administered questionnaire is available for the 36-item, 12-item, and 12 + 24 item WHODASII (see translations/adaptations section for additional information about these versions). The self-administered WHODASII is available in the 36-item and 12-item format. A proxy version of the WHODASII can be completed by others, including clinicians. Also available in 36-item and 6-item versions.

Scoring. Detailed scoring information and updates are available to those who access the WHODASII web site and register their study. In general, total and subscale (i.e., domain) scores are based upon a weighted sum of items and then transformed into a standard scale (0–100). Scores for those working or in school are based on all 36 items. Scores for those not working or in school are based on 32 items (i.e., omitting the work/school questions of the life activities domain). Mean scores can be used to assign a value for missing data. Simple scoring (i.e., summing scores for each domain/no weighting) can be calculated to facilitate use in the clinic, but should not be used to compare with other samples. Complex scoring using item response theory is available on the WHO web site.

Score interpretation. Transformed scores range from 0–100 with higher scores indicating greater disability (i.e., more activity limitations and participation restrictions).

Respondent burden. The 36-item version takes ~20 minutes to complete. The 12-item version takes ~5 minutes. Written and verbal prompts are provided to help respondents, and the interviewer-administered version can aid participants with literacy and other difficulties completing the questionnaire. A proxy version is also available. The questionnaire is not burdensome in terms of reading level required or emotional content.

Administrative burden. Some training is required for the interviewer-administered questionnaire. A detailed training manual is available that facilitates training. Scoring difficulty is moderate.

Translations/adaptations. A 12-item brief assessment of the WHODASII includes 1 or 2 items from each of the domains in the 36-item version plus 3 global questions asking about number of days with health difficulties in past 30 days, number of days of being unable to carry out usual activities, and number of days for which activities were cut back or reduced. A 12 + 24-item screener is available in an interviewer-administered format. This version asks 12 items to screen for disability. If respondents indicate any difficulties, they are asked up to 24 additional questions according to the interviewer guide. Languages available include Albanian, Arabic, Bengali, Chinese (Mandarin), Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hindi, Italian, Japanese, Kannada, Korean, Norwegian, Portuguese, Romanian, Russian, Serbian, Slovenian, Spanish, Sinhala, Swedish, Tamil, Thai, Turkish, and Yoruba. The WHO is active in updating and adapting the WHODASII.

Psychometric Information

Method of development. The WHODASII was developed in collaboration with the WHO, National Institute on Mental Health, National Institute on Alcohol Abuse and Alcoholism, and National Institute on Drug Abuse. Collectively, they make up the WHO Classification, Terminology and Standards team. Development included a 19-country cross-cultural sample for psychometric analysis and screening. Field testing occurred in 2 waves and included members of the general population in good health, people with physical disorders/conditions, people with mental or emotional disorders, and people with problems related to alcohol or drug use (104,105). Psychometric analyses included confirmatory factor analysis, nonparametric, and parametric methods of item response theory testing (104,106–108).

Acceptability. Overall, missing data are low. However, questions related to employment, school, and sexual activities have higher amounts of missing data or refusal rates. Floor effects (i.e., no problems) have been found most frequently in the domains measuring self-care and getting along with others (106–108).

Reliability. Internal consistency of the total or global WHODASII using Cronbach's alpha coefficients is often in the range of 0.86–0.95 for the interviewer-administered and self-administered 36-item versions (102,105–108). Internal consistency across the different domains of the WHODASII often exceed 0.85, although Cronbach's alpha levels have sometimes been much lower for the following domains: getting along with others (often below 0.75) and self-care (0.69–0.82) (102,107,108). Test–retest intraclass correlations across domains and using the total scale have typically been high (e.g., 0.82–0.96 in a sample of individuals with scleroderma when administered after 1 week [102]; 0.87–0.97 in a sample of individuals with knee OA [107]).

Validity. The WHODASII has undergone extensive psychometric analyses including confirmatory factor analysis, nonparametric, and parametric methods of item response theory testing (104,106–109) in samples of individuals with diverse health conditions, including rheumatic diseases and musculoskeletal disorders. Data for arthritis has often been combined with those of other diseases and not presented separately. However, criterion, construct, and discriminant validity for the WHODASII in samples with different types of arthritis have been good to excellent (102,103,105–111). For example, domain and total scores have significantly correlated with clinical disease features (e.g., tender and swollen joint counts, pain, fatigue), other measures of disability and functioning (e.g., Short Form 36, Disease Activity Score in 28 joints, Health Assessment Questionnaire disability index, Western Ontario and McMaster Universities OA Index, Nottingham Health Profile), and have discriminated between pain and disease severity groups (102,103,107–111).

Ability to detect change. Research by the WHO examining the predictive validity and sensitivity to change of the WHODASII is ongoing. Research examining responsiveness and sensitivity to change in samples with arthritis is lacking. However, in a study examining a rehabilitation intervention with different disease groups that included RA and OA, there were small to modest effect sizes ranging from 0.16–0.69 found (108). Similar effect sizes were found in a 3-week spa intervention with individuals with ankylosing spondylitis (103).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Emerging psychometric work using WHODASII across a range of arthritis diagnoses is promising. WHODASII is actively being tested in a variety of countries making its use in international, comparative studies promising. The measure captures a range of elements important to arthritis.

Caveats and cautions. More data are needed using samples of patients with arthritis, especially to examine responsiveness to change. Because not all domains of the WHODASII are equally relevant to all diseases, there may be floor effects in some areas (e.g., cognition).

Clinical usability. Additional data are needed to support use of WHODASII in clinical and treatment settings. The wide range of versions available (e.g., 36 item, 12 item) suggest that administrative burden in the clinic should not be a problem.

Research usability. Initial psychometric evaluation of the WHODASII and its potential for use in international, comparative studies provides support for its use in research. Additional research with different versions of the questions would be beneficial.

LATE-LIFE FUNCTION AND DISABILITY INSTRUMENT (LLFDI)

Description

Purpose. The LLFDI is a general measure of physical disablement developed for older adults. It can be used across a wide range of health levels and conditions and has been applied to samples of middle-aged and even younger adults. First published in 2002, the LLFDI was revised to create an abbreviated version in 2005 and a computer-adapted version in 2008 (112–115). The abbreviated version of the LLFDI is presented elsewhere in this article.

Content. Physical disablement is measured with questions about personal maintenance; mobility and travel; exchange of information; social, community, and civic activities; home life; paid or volunteer work; and involvement in economic activities. These are divided into 2 components: function (difficulty with basic tasks involving lower-extremity function [e.g., walking, climbing stairs, sitting down and standing up, running short and longer distances, getting in and out of a car/taxi] and upper-extremity function [e.g., using utensils, reaching overhead, putting on and taking off a coat or jacket]) and disability (measures the frequency as well as limitations in activities and roles like visiting friends and family, providing care to others, volunteer work, household chores, fitness activities, errands, and personal care needs).

Number of items. In total, the LLFDI has 64 items. Eight additional questions are asked of individuals who use a cane, walker, or other walking device, bringing the total to 72 questions. Within the function subscale, there are questions assessing upper-extremity function (7 items), basic lower-extremity function (14 items), and advanced lower-extremity function (11 items). The disability subscale assesses 16 activities/roles. For each activity/role, respondents are asked to indicate how frequently they perform the activity and to what extent they feel limited in their performance. Frequency questions assess social roles (9 items) and personal roles (7 items), and limitation questions assess instrumental roles (12 items) and management roles (4 items).

Response options/scale. Function questions ask about difficulty with tasks and are measured on a 5-point Likert-type scale, where 1 = cannot do, 2 = quite a lot, 3 = some, 4 = a little, and 5 = none. Disability questions ask “how often do you . . .” with responses on a 5-point Likert-type scale, where 1 = never, 2 = almost never, 3 = once in awhile, 4 = often, and 5 = very often. Activities are also responded to in terms of “to what extent do you feel limited in . . .” with responses on a 5-point scale, where 1 = completely, 2 = a lot, 3 = somewhat, 4 = a little, and 5 = not at all.

Recall period for items. A typical or average day.

Endorsements. There are no known endorsements.

Examples of use. The LLFDI has been used with older adults across a range of health conditions including osteoarthritis, multiple sclerosis, stroke, heart disease, cancer, urinary incontinence, rehabilitation studies, and general population studies of older adults. The measure has also been used as a screening tool, to describe the impact of various health conditions, and as an outcome in intervention studies such as physical activity and health care service interventions (116–121).

Practical Application

How to obtain. The LLFDI can be used with permission by consulting its developer, Alan Jette, e-mail: ajette@bu.edu. Copyright is held by Boston University. Information on costs was unavailable.

Method of administration. The original LLFDI was developed as an interviewer-administered questionnaire (113). However, self-administered (122,123) and telephone-administered formats have also been used and a computerized adaptive test (CAT) has been developed (114).

Scoring. Total scores can be obtained for both the function and disability components. Items are summed and then transformed to create a score ranging from 0–100, where 0 indicates poor ability (i.e., greater difficulty and more limitations) and infrequent performance, and 100 indicates good function and ability (i.e., less difficulty and fewer limitations) and frequent activity performance. Computer scoring is available. Some alternative scoring options exist for the function and disability components of the LLFDI and were based upon factor analyses and Rasch scaling techniques (113). Factor analyses of the items in the function component yielded 3 subscales measuring upper-extremity function (7 items; e.g., reaching, holding a glass, using utensils), basic lower-extremity function (14 items; e.g., climbing stairs, bending overhead, making a bed, getting on/off a bus), and advanced lower-extremity function (11 items; e.g., carrying while climbing stairs, hiking, getting up off the floor, walking a brisk mile, running to catch a bus). Factor analyses of the disability component measuring frequency of activities yielded 2 subscales measuring social roles (9 items; e.g., inviting family and friends into home, traveling out of town, keeping in touch with others, going out to public places, active recreation) and personal roles (7 items; e.g., errands, meal preparation, personal care needs, taking care of household business). Factor analyses of limitation items also yielded 2 factors of different items tapping instrumental roles (12 items; e.g., taking care of the inside of home, errands, socializing, meal preparation) and management roles (4 items; e.g., taking care of health, taking care of household business, keeping in touch with others). Rasch scaling analyses supported both the 1- and 2-factor solutions as reasonable hierarchical scales.

Score interpretation. 0–100, where 0 indicates poor ability (i.e., greater difficulty and more limitations) and infrequent performance, and 100 indicates good function and ability (i.e., less difficulty and fewer limitations) and frequent activity performance. Normative values are not available.

Respondent burden. Takes ~20–25 minutes to complete. The LLFDI is a lengthy questionnaire, but not burdensome in terms of reading level required or emotional content. The computer-assisted version is considerably quicker and algorithms enable items to be skipped based on answers to previous questions.

Administrative burden. Some interviewer training is needed for interviewer-administered questionnaires. Scoring is relatively simple.

Translations/adaptations. Available in English, German, and Hebrew.

Psychometric Information

Method of development. Items were generated and refined based on a review of the literature, consultation with experts, and input from community focus groups with older adults. Further refinements of the LLFDI were made and subscales developed using exploratory factor analysis and Rasch analysis (112,113).

Acceptability. The LLFDI is relatively easy to complete. To date, studies report no or minimal floor and ceiling effects. Interviewer-administered questionnaires have resulted in little in the way of missing data (112,113,123–125).

Reliability. For the function component of the LLFDI, evidence for test–retest reliability over 1–3 weeks (average 12 days) has been very good across different samples of older adults. Intraclass correlations have ranged from 0.77–0.98. Internal consistency, using Cronbach's alpha, for the 3 subscales of advanced lower-extremity function, basic lower-extremity function, and upper-extremity function were 0.96, 0.96, and 0.86, respectively. A combined function scale had a Cronbach's alpha of 0.97 (112,120,124). For the disability component of the LLFDI, test–retest intraclass correlations across 1–3 weeks (average 12 days) have been modest to good, ranging from 0.63–0.83 (113,120,124). Internal consistency of the disability component subscale measuring the frequency of activities was 0.82 using Cronbach's alpha. Internal consistency of the disability component subscale measuring limitations in activities was 0.92 using Cronbach's alpha (n = 150) (113).

Validity. Research is needed to examine the validity of the LLFDI in samples with arthritis. However, results are promising in samples of older adults with a range of chronic health conditions. Psychometric analyses have compared scores on the function and disability components to performance tests such as 400-Meter Walk Test, Short Physical Performance Battery, 2-Minute Walk Distance, 8-Foot Walk Test, Berg Balance Scale, and Timed Up & Go Test, as well as self-report questionnaires like the physical functioning scale and physical component of the Short Form 36. (120,122,123,125,126). Results found that the LLFDI demonstrated concurrent and predictive validity with functional performance using 400-Meter Walk Test and Short Physical Performance Battery. The function component of the LLFDI demonstrated substantial associations with functional performance measures, which were strongest for the overall and lower-extremity function dimensions. With respect to predictive validity, it was found that performance measures of physical function predicted limitations in daily activities in the disability component of the LLFDI (123). Cross-sectional convergent validity of the LLFDI also was supported when applied to adults 45–65 years of age with chronic conditions (127).

Ability to detect change. The time frame for the LLFDI is a typical or average day. Asking participants to characterize their general disability may result in difficulties in the measure's ability to detect small changes. Research also needs to be conducted in arthritis. To date, studies with older adults have used the LLFDI as an outcome in research on the use of antidepressants (118,122) and a physical activity intervention for stroke patients (121) with the LLFDI showing significant changes as a result of treatment and intervention. Olarsch (128) reported that the LLFDI was more responsive than the EQ-5D and Elderly Mobility Scale in a group of older adults living in long-term care.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The LLFDI covers a range of disability domains not found in many arthritis-specific measures, including a participation in diverse social activities and roles. It also asks participants for information about the frequency of involvement in activities and roles, as well as limitations. As a result, it is a fairly comprehensive measure with good potential for use with older adults who have a range of chronic health conditions, including arthritis.

Caveats and cautions. Currently, there is no detailed psychometric evaluation of the measure with arthritis samples. The length of the measure may limit its use, although a CAT version of the LLFDI may enable quicker administration in some settings.

Clinical usability. Although psychometric evaluation supports the use of the LLFDI in samples of older adults in long-term care, depressed older adults, and stroke patients, clinical research in arthritis is needed. The length of the LLFDI may be a barrier to its clinical use.

Research usability. More research is needed in arthritis, but the LLFDI is a potentially attractive tool as its items include both activity limitations and disability with roles. The length of the LLFDI may be a barrier to its research use.

LATE-LIFE FUNCTION AND DISABILITY INSTRUMENT-ABBREVIATED VERSION (LLFDI-ABBREVIATED)

Description

Purpose. The LLFDI-Abbreviated is a general measure of physical disablement developed for older adults. It can be used across a wide range of health levels and conditions. The LLFDI-Abbreviated was created in 2005 (115).

Content. The 2 components of the original LLFDI were maintained with disablement measured with questions about function and disability. The function component maintained separate subscales related to upper-extremity function (e.g., holding a glass of water, using utensils, unscrewing a lid), basic lower-extremity function (e.g., getting in/out of a car, bending over while standing, walking around the home), and advanced lower-extremity function (e.g., carrying while climbing stairs, walking 1 mile with rests, going up/down 3 flights of stairs). Disability component items measuring frequency and limitations in activities have subscales for social (e.g., go out to public places, visit friends) and personal roles (e.g., errands, household business). It is worth noting here that the original LLFDI labeled factors derived from disability component limitation items as instrumental roles and management roles.

Number of items. The total LLFDI-Abbreviated has 31 items. The function component has retained 15 of the original 32 items measuring upper-extremity (5 items), basic lower-extremity (5 items), and advanced lower-extremity function (5 items). The disability component retained 8 of the original 16 items (social roles = 4 items, personal roles = 4 items).

Response options/scale. Function questions ask about difficulty with tasks and are measured on a 5-point Likert-type scale, where 1 = cannot do, 2 = quite a lot, 3 = some, 4 = a little, and 5 = none. Disability questions ask about frequency of activities (i.e., "how often do you . . .") with responses on a 5-point Likert-type scale, where 1 = never, 2 = almost never, 3 = once in awhile, 4 = often, and 5 = very often. Activities are also responded to in terms of limitations (i.e., "to what extent do you feel limited in . . .") with responses on a 5-point scale, where 1 = completely, 2 = a lot, 3 = somewhat, 4 = a little, and 5 = not at all.

Recall period for items. A typical day or average day.

Endorsements. There are no known endorsements.

Examples of use. Community samples of adults (age ≥60), geriatric inpatients, older adults, and physical activity changes (115,129–132).

Practical Application

How to obtain. See information for original LLFDI. Items in the LLFDI-Abbreviated are outlined in reference (115).

Method of administration. Psychometric testing has mostly used an interviewer-administered LLFDI-Abbreviated (115). However, a mailed, self-administered questionnaire has also been used in research (132,133).

Scoring. Scoring practices differ from the original LLFDI. The abbreviated version sums item scores across the function and disability subscales (115,131). Total scores for the function and disability components can also be calculated.

Score interpretation. Function component subscales (i.e., upper extremity, basic lower extremity, advanced lower extremity) range from 5–25. Total scores range from 15–75. Higher scores indicate fewer functional limitations (i.e., less disability). Disability component subscales (i.e., social and personal roles) range from 4–20. Total scores range from 8–40. Higher scores indicate less disability. No cut points or normative values have been established.

Respondent burden. Time to complete ranged from under 10 minutes to ~30 minutes in a sample of older community-dwelling adults (129). Questions are not burdensome in terms of reading level required or emotional content.

Administrative burden. Some interviewer training is needed for interviewer-administered questionnaires. Scoring is relatively simple.

Translations/adaptations. Available in English and German.

Psychometric Information

Method of development. A reduced number of items were selected from the original LLFDI using confirmatory factor analysis with maximum likelihood estimation in LISREL software used to establish model fit parameters (115).

Acceptability. Denkinger and colleagues report minimal floor and ceiling effects in the German LLFDI-Abbreviated and a good range of scores (129,130).

Reliability. Test–retest reliability for the LLFDI-Abbreviated function component was very good (0.81–0.96) and interrater reliability was acceptable to very good (0.62–0.96) with the German LLFDI-Abbreviated (129). No information about test–retest reliability was available for the disability component. Internal consistency as measured by Cronbach's alpha ranged from: upper-extremity function 0.58–0.84; basic lower-extremity function 0.76–0.83; and advanced lower-extremity function 0.80–0.86 (115,129,130). Cronbach's alphas for the disability component subscales of the LLFDI-Abbreviated were lower, especially for items asking about the frequency of social role activities (alphas ranged from 0.38–0.67). Cronbach's alphas for the personal role subscales were better (alphas ranged from 0.77–0.83). Items asking about limitations in social and personal roles ranged from 0.77–0.83 (115).

Validity. The original LLFDI and LLFDI-Abbreviated were moderately to highly correlated with one another. The relationship between the function component subscales on the 2 versions ranged from 0.92–0.97 and the relationship between the 2 versions of the disability component subscales ranged from 0.76–0.80 (115). The LLFDI-Abbreviated subscales were significantly associated with performance tests such as the Physical Activity Scale for the Elderly and the Community Healthy Activities Model Program for Seniors. Correlations were typically greater for the function component subscales compared to the disability component (115). In a sample of 292 adults with multiple sclerosis ranging in age from 20–69 years (mean age 48 years), LLFDI-Abbreviated components were moderately to highly related to neurological impairments and symptoms, as well as poorer health status and quality of life in the expected directions (133). The function component score also was found to be moderately to highly correlated with the Fall Efficacy Scale International, the Short Physical Performance Battery, the Timed Up & Go Test, and other performance-based measures (e.g. normal speed, maximum speed, step length; Spearman's correlations ranged from 0.42–0.76) (130).

Ability to detect change. Data are limited. However, Denkinger and colleagues used standardized response mean (SRM) values to evaluate sensitivity to change across a 3-week period for the function component of the LLFDI-Abbreviated. SRM values were all significant with medium effect sizes, varying with the treatment period (130).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Similar to the original version, the LLFDI-Abbreviated covers a range of disability domains not found in many arthritis-specific measures, including a participation in diverse social activities and roles. It asks participants for information about the frequency of involvement in activities and roles, as well as limitations. Its shorter length may make it useful in studies with individuals who have arthritis. Preliminary findings suggest that the measure may also be applicable to younger and middle-aged samples.

Caveats and cautions. Additional psychometric analyses of the measure are needed to assess reliability, validity, and sensitivity to change.

Clinical usability. More research is needed prior to a recommendation on the clinical usability of the LLFDI-Abbreviated.

Research usability. More research is needed in arthritis, but the LLFDI-Abbreviated is a potentially attractive tool as its items include both activity limitations and disability with roles. The shorter length of this version may make it more feasible for use.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. World Health Organization. International classification of functioning, disability, and health. Geneva: World Health Orgaization; 2001.
2. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2: the content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. Arthritis Rheum 1992;35:1–10.
3. Guillemin F, Coste J, Pouchot J, Ghezail M, Bregeon C, Sany J, and the French Quality of Life in Rheumatology Group. The AIMS2-SF: a short form of the Arthritis Impact Measurement Scales 2. Arthritis Rheum 1997;40:1267–74.
4. Hagen KB, Smedstad LM, Uhlig T, Kvien TK. The responsiveness of health status measures in patients with rheumatoid arthritis: comparison of disease-specific and generic instruments. J Rheumatol 1999; 26:1474–80.
5. Hakala M, Nieminen P, Manelius J. Joint impairment is strongly correlated with disability measured by self-reported questionnaires: functional status assessment of individuals with rheumatoid arthritis in a population based series. J Rheumatol 1994;21:64–9.
6. Salaffi F, Stancati A, Carotti M. Responsiveness of health status measures and utility-based methods in patients with rheumatoid arthritis. Clin Rheumatol 2002;21:478–87.
7. Ten Klooster PM, Veehof MM, Taal E, van Riel PL, van de Laar MA. Changes in priorities for improvement in patients with rheumatoid arthritis during 1 year of anti-tumour necrosis factor treatment. Ann Rheum Dis 2007;66:1485–90.
8. Rosemann T, Joos S, Laux G, Gensichen J, Szecsenyi J. Case management of arthritis patients in primary care: a cluster-randomized controlled trial. Arthritis Rheum 2007;57:1390–7.
9. Husted J, Gladman DD, Farewell VT, Long JA. Validation of the revised and expanded version of the Arthritis Impact Measurement Scales for patients with psoriatic arthritis. J Rheumatol 1996;23: 1015–9.
10. Husted J, Gladman DD, Long JA, Farewell VT. Relationship of the Arthritis Impact Measurements Scales (AIMS) to changes in articular status and functional performance in patients with psoriatic arthritis (PsA). J Rheumatol 1996;23:1923–37.
11. Long JA, Husted JA, Gladman DD, Farewell VT. The relationship between patient satisfaction with health and clinical measures of

function and disease status in patients with psoriatic arthritis. J Rheumatol 2000;27:958–66.

12. Taccari E, Spadaro A, Rinaldi T, Riccieri V, Sensi F. Comparison of the Health Assessment Questionnaire and Arthritis Impact Measurement Scale in patients with psoriatic arthritis. Rev Rhum Eng Ed 1998;65:751–8.
13. Guillemin F, Challier B, Urlacher F, Vancon G, Pourel J. Quality of life in ankylosing spondylitis: validation of the ankylosing spondylitis Arthritis Impact Measurement Scales 2, a modified Arthritis Impact Measurement Scales Questionnaire. Arthritis Care Res 1999;12:157–62.
14. Challier B, Urlacher F, Vancon G, Lemelle I, Pourel J, Guillemin F. Is quality of life affected by season and weather conditions in ankylosing spondylitis? Clin Exp Rheumatol 2001;19:277–81.
15. Lukaczer D, Darland G, Tripp M, Liska D, Lerman RH, Schiltz B, et al. A pilot trial evaluating Meta050, a proprietary combination of reduced iso-alpha acids, rosemary extract and oleanolic acid in patients with arthritis and fibromyalgia. Phytother Res 2005;19:864–9.
16. Neumann L, Dudnik Y, Bolotin A, Buskila D. Evaluation of a Hebrew version of the revised and expanded Arthritis Impact Measurement Scales (AIMS2) in patients with fibromyalgia. J Rheumatol 1999;26:1816–21.
17. Amadio PC, Silverstein MD, Ilstrup DM, Schleck CD, Jensen LM. Outcome after Colles fracture: the relative responsiveness of three questionnaires and physical examination measures. J Hand Surg Am 1996;21:781–7.
18. Beeton K. Evaluation of outcome of care in patients with haemophilia. Haemophilia 2002;8:428–34.
19. De Joode EW, van Meeteren NL, van den Berg HM, de Kleijn P, Helders PJ. Validity of health status measurement with the Dutch Arthritis Impact Measurement Scale 2 in individuals with severe haemophilia. Haemophilia 2001;7:190–7.
20. Meeteren van NL, Strato IH, van Veldhoven NH, De Kleijn P, van den Berg HM, Helders PJ. The utility of the Dutch Arthritis Impact Measurements Scales 2 for assessing health status with haemophilia; a pilot study. Haemophilia 2000;6:664–71.
21. Poiraudeau S, Dougados M, Ait-Hadad H, Pion-Graff J, Ayral X, Listrat V, et al. Evaluation of a quality of life scale (AIMS2) in rheumatology. Rev Rhum Ed Fr 1993;60:561–7.
22. McDowell I. Measuring health: A guide to rating scales and questionnaires 2006. New York: Oxford University Press.
23. Brandao L, Ferraz MB, Zerbini CA. Health status in rheumatoid arthritis: cross cultural evaluation of a Portuguese version of the Arthritis Impact Measurement Scales 2 (BRASIL-AIMS2). J Rheumatol 1998;25:1499–501.
24. Chu EM, Chiu KY, Wong RW, Tang WM, Lau CS. Translation and validation of Arthritis Impact Measurement Scales 2 into Chinese: CAIMS2. Arthritis Rheum 2004;51:20–7.
25. Mousavi SJ, Parnianpour M, Askary-Ashtiani AR, Hadian MR, Rostamian A, Montazeri A. Translation and validation study of the Persian version of the Arthritis Impact Measurement Scales 2 (AIMS2) in patients with osteoarthritis of the knee. BMC Musculoskelet Disord 2009;10:95.
26. Rosemann T, Szecsenyi J. Cultural adaptation and validation of a German version of the Arthritis Impact Measurement Scales (AIMS2). Osteoarthritis Cartilage 2007;15:1128–33.
27. Spies-Dorgelo MN, Terwee CB, Stalman WA, van der Windt DA. Reproducibility and responsiveness of the Symptom Severity Scale and the hand and finger function subscale of the Dutch arthritis impact measurement scales (Dutch-AIMS2-HFF) in primary care patients with wrist or hand problems. Health Qual Life Outcomes 2006;4:87.
28. Coulton CJ, Zborowsky E, Lipton J, Newman A. Assessment of the reliability and validity of the arthritis impact measurement scales for children with juvenile arthritis. Arthritis Rheum 1987;30:819–24.
29. Hughes SL, Edelman P, Chang RW, Singer RH, Schuette P. The GERI-AIMS: reliability and validity of the arthritis impact measurement scales adapted for elderly respondents. Arthritis Rheum 1991;34:856–65.
30. Atamaz F, Hepguler S, Oncu J. Translation and validation of the Turkish version of the arthritis impact measurement scales 2 in patients with knee osteoarthritis. J Rheumatol 2005;32:1331–6.
31. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992;30:917–25.
32. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis: the Arthritis Impact Measurement Scales. Arthritis Rheum 1980;23:146–52.
33. Meenan RF, Gertman PM, Mason JH, Dunaif R. The Arthritis Impact Measurement Scales: further investigations of a health status measure. Arthritis Rheum 1982;25:1048–53.
34. Meenan RF. New approaches to outcome assessment: the AIMS questionnaire for arthritis. Adv Intern Med 1986;31:167–85.
35. Haavardsholm EA, Kvien TK, Uhlig T, Smedstad LM, Guillemin F. A comparison of agreement and sensitivity to change between AIMS2 and a short form of AIMS2 (AIMS2-SF) in more than 1,000 rheumatoid arthritis patients. J Rheumatol 2000;27:2810–6.
36. Ward MM, Kuzis S. Validity and sensitivity to change of spondylitis-specific measures of functional disability. J Rheumatol 1999;26:121–7.
37. Taal E, Rasker JJ, Riemsma RP. Sensitivity to change of AIMS2 and AIMS2-SF components in comparison to M-HAQ and VAS-pain. Ann Rheum Dis 2004;63:1655–8.
38. Husted JA, Gladman DD, Cook RJ, Farewell VT. Responsiveness of health status instruments to changes in articular status and perceived health in patients with psoriatic arthritis. J Rheumatol 1998;25:2146–55.
39. Baillet A, Payraud E, Niderprim VA, Nissen MJ, Allenet B, Francois P, et al. A dynamic exercise programme to improve patients' disability in rheumatoid arthritis: a prospective randomized controlled trial. Rheumatology (Oxford) 2009;48:410–5.
40. Taylor LF, Kee CC, King SV, Ford TA. Evaluating the effects of an educational symposium on knowledge, impact, and self-management of older African Americans living with osteoarthritis. J Community Health Nurs 2004;21:229–38.
41. Wetzels R, van Weel C, Grol R, Wensing M. Family practice nurses supporting self-management in older patients with mild osteoarthritis: a randomized trial. BMC Fam Pract 2008;9:7.
42. Ren XS, Kazis L, Meenan RF. Short-form Arthritis Impact Measurement Scales 2: tests of reliability and validity among patients with osteoarthritis. Arthritis Care Res 1999;12:163–71.
43. Ten Klooster PM, Veehof MM, Taal E, Van Riel PL, Van de Lar MA. Confirmatory factor analysis of the Arthritis Impact Measurement Scales 2 Short Form in patients with rheumatoid arthritis. Arthritis Rheum 2008;59:692–8.
44. Rosemann T, Korner T, Wensing M, Schneider A, Szecsenyi J. Evaluation and cultural adaptation of a German version of the AIMS2-SF questionnaire (German AIMS2-SF). Rheumatology (Oxford) 2005;44:1190–5.
45. Taal E, Rasker JJ, Riemsma RP. Psychometric properties of a Dutch short form of the Arthritis Impact Measurement Scales 2 (Dutch-AIMS2-SF). Rheumatology (Oxford) 2003;42:427–34.
46. Rat AC, Guillemin F, Pouchot J. Mapping the osteoarthritis knee and hip quality of life (OAKHQOL) instrument to the international classification of functioning, disability and health and comparison to five health status instruments used in osteoarthritis. Rheumatology (Oxford) 2008;47:1719–25.
47. Askary-Ashtiani AR, Mousavi SJ, Parnianpour M, Montazeri A. Translation and validation of the Persian version of the Arthritis Impact Measurement Scales 2-Short Form (AIMS2-SF) in patients with rheumatoid arthritis. Clin Rheumatol 2009;28:521–7.
48. Wijilhuizen GJ, Ooijendijk W. Measuring disability, the agreement between self evaluation and observation of performance. Disabil Rehabil 1999;21:61–7.
49. McDowell I. Physical disability and handicap: the OECD long term disability questionnaire. In: Measuring health. A guide to ratings scales and questionnaires. New York: Oxford University Press; 2006. p. 93–5.
50. Wilson RW, McNeil JM. Preliminary analysis of the OECD disability on the pretest of the post census disability survey. Rev Epidemiol Sante Publique 1981;29:469–75.
51. Badley EM, Rothman LM, Wang PP. Modelling physical dependence in arthritis: the relative contribution of specific disabilities and environmental factors. Arthritis Care Res 1998;11:335–49.
52. Kaptein SA, Gignac MA, Badley EM. Differences in the workforce experiences of women and men with arthritis disability: a population health perspective. Arthritis Rheum 2009;61:605–13.
53. Statistics Canada. Report of the Canadian Health and Disability Survey (STC 82-555E/F). Ottawa: Statistics Canada; 1986.
54. Statistics Canada. The 2006 participation and activity limitation survey: disability in Canada (c89-628-XIE). Ottawa: Minister of Industry; 2010.
55. Wang PP, Badley EM. The contribution of arthritis and arthritis disability to non-participation in the labour force: a Canadian example. J Rheumatol 2001;28:1077–82.
56. The EuroQol Group. EuroQol: a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
57. Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H, et al. Validity of Euroqol, a generic health status instrument, in patients with rheumatoid arthritis. Br J Rheumatol 1994;33:655–62.
58. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). Br J Rheumatol 1997;36:551–9.
59. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. Comparing the standard EQ-5D three-level system with a five-level version. Value Health 2008;11:275–84.
60. Fransen M, Edmonds J. Reliability and validity of the EuroQol in

patients with osteoarthritis of the knee. Rheumatology (Oxford) 1999; 38:807–13.

61. Wolfe F, Kong SX, Watson DJ. Gastrointestinal symptoms and health related quality of life in patients with arthritis. J Rheumatol 2000;27: 1373–8.
62. Xie F, Li SC, Luo N, Lo NN, Yeo SJ, Yang KY, et al. Comparison of the EuroQol and short form 6D in Singapore multiethnic Asian knee osteoarthritis patients scheduled for total knee replacement. Arthritis Rheum 2007;57:1043–9.
63. Gordeev VS, Maksymowych WP, Evers SM, Ament A, Schachna L, Boonen A. Role of contextual factors in health-related quality of life in ankylosing spondylitis. Ann Rheum Dis 2010;69:108–12.
64. Haywood KL, Garratt AM, Dziedzic K, Dawes PT. Generic measures of health-related quality of life in ankylosing spondylitis: reliability, validity and responsiveness. Rheumatology (Oxford) 2002;41:1380–7.
65. Khanna D, Ahmed M, Yontz D, Ginsburg SS, Park GS, Leonard A, et al. The disutility of chronic gout. Qual Life Res 2008;17:815–22.
66. Duarte-Salazar C, Guzman-Vazquez S, Soto-Molina H, Chaidez-Rosales P, Ilizaliturri-Sanchez V, Nieves-Silva J, et al. Disability impact on quality of life in Mexican adults with juvenile idiopathic arthritis and juvenile ankylosing spondylitis. Clin Exp Rheumatol 2007;25:922–7.
67. Campbell H, Rivero-Arias O, Johnston K, Gray A, Fairbank J, Frost H. Responsiveness of objective, disease-specific, and generic outcome measures in patients with chronic low back pain: an assessment for improving, stable, and deteriorating patients. Spine (Phila Pa 1976) 2006;31:815–22.
68. Garratt AM, Klaber Moffett J, Farrin AJ. Responsiveness of generic and specific measures of health outcome in low back pain. Spine (Phila Pa 1976) 2001;26:71–7.
69. Aggarwal R, Wilke CT, Pickard AS, Vats V, Mikolaitis R, Fogg L, et al. Psychometric properties of the EuroQol-5D and Short Form-6D in patients with systemic lupus erythematosus. J Rheumatol 2009;36: 1209–16.
70. Wolfe F, Michaud K, Li T, Katz RS. EQ-5D and SF-36 quality of life measures in systemic lupus erythematosus: comparisons with rheumatoid arthritis, noninflammatory rheumatic disorders, and fibromyalgia. J Rheumatol 2010;37:296–304.
71. Mease PJ, Woolley JM, Singh A, Tsuji W, Dunn M, Chiou CF. Patient-reported outcomes in a randomized trial of etanercept in psoriatic arthritis. J Rheumatol 2010;37:1221–7.
72. Kievit W, Adang EM, Fransen J, Kuper HH, van de Laar MA, Jansen TL, et al. The effectiveness and medication costs of three anti-tumour necrosis factor α agents in the treatment of rheumatoid arthritis from prospective clinical practice data. Ann Rheum Dis 2008;67:1229–34.
73. Teng YK, Verburg RJ, Sont JK, van den Hout WB, Breedveld FC, van Laar JM. Long-term followup of health status in patients with severe rheumatoid arthritis after high-dose chemotherapy followed by autologous hematopoietic stem cell transplantation. Arthritis Rheum 2005; 52:2272–6.
74. Van den Hout WB, de Jong Z, Munneke M, Hazes JM, Breedveld FC, Vliet Vlieland TP. Cost-utility and cost-effectiveness analyses of a long-term, high-intensity exercise program compared with conventional physical therapy in patients with rheumatoid arthritis. Arthritis Rheum 2005;53:39–47.
75. Wolfe F, Michaud K. The loss of health status in rheumatoid arthritis and the effect of biologic therapy: a longitudinal observational study. Arthritis Res Ther 2010;12:R35.
76. Pipitone N, Scott DL. Magnetic pulse treatment for knee osteoarthritis: a randomised, double-blind, placebo-controlled study. Curr Med Res Opin 2001;17:190–6.
77. Raman R, Dutta A, Day N, Sharma HK, Shaw CJ, Johnson GV. Efficacy of Hylan G-F 20 and Sodium Hyaluronate in the treatment of osteoarthritis of the knee: a prospective randomized clinical trial. Knee 2008;15:318–24.
78. Boonen A, Patel V, Traina S, Chiou CF, Maetzel A, Tsuji W. Rapid and sustained improvement in health-related quality of life and utility for 72 weeks in patients with ankylosing spondylitis receiving etanercept. J Rheumatol 2008;35:662–7.
79. Braun J, McHugh N, Singh A, Wajdula JS, Sato R. Improvement in patient-reported outcomes for patients with ankylosing spondylitis treated with etanercept 50 mg once-weekly and 25 mg twice-weekly. Rheumatology (Oxford) 2007;46:999–1004.
80. Van Tubergen A, Boonen A, Landewe R, Rutten-Van Molken M, Van Der Heijde D, Hidding A, et al. Cost effectiveness of combined spa–exercise therapy in ankylosing spondylitis: a randomized controlled trial. Arthritis Rheum 2002;47:459–67.
81. Epps H, Ginnelly L, Utley M, Southwood T, Sculpher M, Woo P. Is hydrotherapy cost-effective? A randomised controlled trial of combined hydrotherapy programmes compared with physiotherapy land techniques in children with juvenile idiopathic arthritis. Health Technol Assess 2005;9:iii–iv, ix–x, 1–59.
82. Larsen K, Sorensen OG, Hansen TB, Thomsen PB, Soballe K. Accelerated perioperative care and rehabilitation intervention for hip and knee replacement is effective: a randomized clinical trial involving 87 patients with 3 months of follow-up. Acta Orthop 2008;79:149–59.
83. Rolfson O, Dahlberg LE, Nilsson JA, Malchau H, Garellick G. Variables determining outcome in total hip replacement surgery. J Bone Joint Surg Br 2009;91:157–61.
84. Barton GR, Sach TH, Avery AJ, Doherty M, Jenkinson C, Muir KR. Comparing the performance of the EQ-5D and SF-6D when measuring the benefits of alleviating knee pain. Cost Eff Resour Alloc 2009;7:12.
85. Adams R, Walsh C, Veale D, Bresnihan B, FitzGerald O, Barry M. Understanding the relationship between the EQ-5D, SF-6D, HAQ and disease activity in inflammatory arthritis. Pharmacoeconomics 2010; 28:477–87.
86. Gulfe A, Kristensen LE, Saxne T, Jacobsson LT, Petersson IF, Geborek P. Utility-based outcomes made easy: the number needed per quality-adjusted life year gained. An observational cohort study of tumor necrosis factor blockade in inflammatory arthritis from Southern Sweden. Arthritis Care Res (Hoboken) 2010;62:1399–406.
87. Huang IC, Willke RJ, Atkinson MJ, Lenderking WR, Frangakis C, Wu AW. US and UK versions of the EQ-5D preference weights: does choice of preference weights make a difference? Qual Life Res 2007; 16:1065–72.
88. Joore M, Brunenberg D, Nelemans P, Wouters E, Kuijpers P, Honig A, et al. The impact of differences in EQ-5D and SF-6D utility scores on the acceptability of cost-utility ratios: results across five trial-based cost-utility studies. Value Health 2010;13:222–9.
89. Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life. J Clin Epidemiol 2003;56:317–25.
90. Luo N, Johnson JA, Shaw JW, Coons SJ. A comparison of EQ-5D index scores derived from the US and UK population-based scoring functions. Med Decis Making 2007;27:321–6.
91. Marra CA, Rashidi AA, Guh D, Kopec JA, Abrahamowicz M, Esdaile JM, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res 2005;14:1333–44.
92. McPherson K, Myers J, Taylor WJ, McNaughton HK, Weatherall M. Self-valuation and societal valuations of health state differ with disease severity in chronic and disabling conditions. Med Care 2004;42: 1143–51.
93. Wilke CT, Pickard AS, Walton SM, Moock J, Kohlmann T, Lee TA. Statistical implications of utility weighted and equally weighted HRQL measures: an empirical study. Health Econ 2010;19:101–10.
94. Wolfe F, Michaud K, Wallenstein G. Scale characteristics and mapping accuracy of the US EQ-5D, UK EQ-5D, and SF-6D in patients with rheumatoid arthritis. J Rheumatol 2010;37:1615–25.
95. Ravens-Sieberer U, Wille N, Badia X, Bonsel G, Burstrom K, Cavrini G, et al. Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study. Qual Life Res 2010;19:887–97.
96. Luo N, Chew LH, Fong KY, Koh DR, Ng SC, Yoon KH, et al. Validity and reliability of the EQ-5D self-report questionnaire in English-speaking Asian patients with rheumatic diseases in Singapore. Qual Life Res 2003;12:87–92.
97. McPhail S, Lane P, Russell T, Brauer SG, Urry S, Jasiewicz J, et al. Telephone reliability of the Frenchay Activity Index and EQ-5D amongst older adults. Health Qual Life Outcomes 2009;7:48.
98. Brodszky V, Pentek M, Balint PV, Geher P, Hajdu O, Hodinka L, et al. Comparison of the Psoriatic Arthritis Quality of Life (PsAQoL) questionnaire, the functional status (HAQ) and utility (EQ-5D) measures in psoriatic arthritis: results from a cross-sectional survey. Scand J Rheumatol 2010;39:303–9.
99. Johnson JA, Pickard AS. Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada. Med Care 2000;38:115–21.
100. Barton GR, Sach TH, Jenkinson C, Avery AJ, Doherty M, Muir KR. Do estimates of cost-utility based on the EQ-5D differ from those based on the mapping of utility scores? Health Qual Life Outcomes 2008;6:51.
101. Kvamme MK, Kristiansen IS, Lie E, Kvien TK. Identification of cutpoints for acceptable health status and important improvement in patient-reported outcomes, in rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis. J Rheumatol 2010;37:26–31.
102. Baron M, Schieir O, Hudson M, Steele R, Kolahi S, Berkson L, et al. The clinimetric properties of the World Health Organization disability assessment schedule II in early inflammatory arthritis. Arthritis Rheum 2008;59:382–90.
103. Van Tubergen A, Landewe R, Heuft-Dorenbosch L, Spoorenberg A, Van Der Heijde D, Van Der Tempel H, et al. Assessment of disability with the World Health Organisation Disability Assessment Schedule II in patients with ankylosing spondylitis. Ann Rheum Dis 2003;62: 140–5.
104. Rehm J, Ustun TB, Shekhar S, Nelson CB, Chatterji S, Ivis F, et al. On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. Int J Methods Psychiatr Res 1999;8:110–22.

105. Ustun TB, Chatterji S, Kostanjsek N, Rehm J, Kennedy C, Epping-Jordan J, et al. Developing the World Health Organization Disability Assessment Schedule 2.0. Bull World Health Organ 2010;88:797–876.
106. Garin O, Ayuso-Mateos JL, Almansa J, Nieto M, Chatterji S, Vilagut G, et al. Validation of the "World Health Organiztion Disablility Assessment Scehdule, WHODAS-2" in patients with chronic diseases. Health Qual Life Outcomes 2010;8:51.
107. Kutlay S, Kucukdeveci AA, Elhan AH, Oztuna D, Koc N, Tennant A. Validation of the World Health Organization disability assessment schedule II (WHODAS-II) in patients with osteoarthritis. Rheumatol Int 2011;31:339–46.
108. Posl M, Cieza A, Stucki G. Psychometric properties of the WHODASII in rehabilitation patients. Qual Life Res 2007;16:1521–31.
109. Meesters JL, Verhoef J, Liem IS, Putter H, Vlieland TP. Validity and responsiveness of the World Health Organization Disability Assessment Schedule II to assess disability in rheumatoid arthritis patients. Rheumatology (Oxford) 2010;49:326–33.
110. Hudson M, Steele R, Taillefer S, Baron M, and the Canadian Scleroderma Research Group. Quality of life in systemic sclerosis: psychometric properties of the World Health Organization Disability Assessment Schedule II. Arthritis Rheum 2008;59:270–8.
111. Noonan VK, Kopec JA, Noreau L, Singer J, Chan A, Masse LC, et al. Comparing the content of participation instruments using the International Classification of Functioning, Disability and Health. Health Qual Life Outcomes 2009;7:93.
112. Haley SM, Jette AM, Coster WJ, Kooyoomjian JT, Levenson S, Heeren T, et al. Late-Life Function and Disability Instrument II: development and evaluation of the function component. J Gerontol A Biol Sci Med Sci 2002;57A:M217–22.
113. Jette AM, Haley SM, Coster WJ, Kooyoomjian JT, Levenson S, Heeren T, et al. Late-Life Function and Disability Instrument I: development and evaluation of the disability component. J Gerontol A Biol Sci Med Sci 2002;57A:M209–16.
114. Jette AM, Haley SM, Ni P, Olarsch S, Moed R. Creating a computer adaptive test version of the late-life function and disability instrument. J Gerontol A Biol Sci Med Sci 2008;63:1246–56.
115. McAuley E, Konopack JF, Motl RW, Rosengren K, Morriset KS. Measuring disability and function in older women: psychometric properties of the Late-Life Function and Disability Instrument. J Gerontol A Biol Sci Med Sci 2005;60A:901–9.
116. Cutchin MP, Coppola S, Talley V, Svihula J, Catellie D, Shank KH. Feasibility and effects of preventive home visits for at-risk older people: design of a randomized controlled trial. BMC Geriatrics 2009; 9:54.
117. Gibson K, Day L, Hill KD, Jolley D, Newstead S, Cicuttini F, et al. Screening for pre-clinical disability in different residential settings. BMC Geriatr 2010;10:52.
118. Lenze EJ, Rollman BL, Shear MK, Dew MA, Pollock BG, Ciliberti C, et al. Escitalopram for older adults with generalized anxiety disorder (GAD): a randomized controlled trial. JAMA 2009;301:295–303.
119. Lowe SS, Watanabe SM, Baracos VE, Courneya KS. Associations between physical activity and quality of life in cancer patients receiving palliative care: a pilot survey. J Pain Symptom Manage 2009;38: 785–96.
120. Melzer I, Marx R, Kurz I. Regular exercise in the elderly is effective to preserve the speed of voluntary stepping under single-task condition but not under dual-task condition: a case-control study. Gerontology 2009;55:49–57.
121. Ouellette MM, LeBrasseur NK, Bean JF, Phillips E, Stein J, Frontera WR, et al. High-intensity resistance training improves muscle strength, self-reported function, and disability in long-term stroke survivors. Stroke 2004;35:1404–9.
122. Karp JF, Skidmore E, Lotz M, Lenze E, Dew MA, Reynolds CF 3rd. Use of the late-life function and disability instrument to assess disability in major depression. J Am Geriatr Soc 2009;57:1612–9.
123. Sayers SP, Jette AM, Haley SM, Heeren T, Guralnik JM, Fielding RA. Validation of the Late-Life Function and Disability Instrument. J Am Geriatr Soc 2004;52:1554–9.
124. LaPier TK, Mizner R. Outcome measures in cardiopulmonary physical therapy: focus on the Late-Life Function and Disability Instrument (LLFDI). Cardiopulm Phys Ther J 2009;20:32–5.
125. Dubuc N, Haley S, Ni P, Kooyoomjian J, Jette A. Function and disability in late life: comparison of the Late-Life Function and Disability Instrument to the Short-Form-36 and the London Handicap Scale. Disabil Rehabil 2004;26:362–70.
126. Melzer I, Kurz I, Sarid O, Jette AM. Relationship between self-reported function and disability and balance performance measures in the elderly. J Rehabil Res Dev 2007;44:685–91.
127. Hand C, Richardson J, Letts L, Stratford P. Construct validity of the Late-Life Function and Disability instrument for adults with chronic conditions. Disabil Rehabil 2010;32:50–6.
128. Olarsch S. Validity and responsiveness of the late-life function and disability instrument (LLFDI) in a facility-dwelling population [dissertation]. Boston: Boston University; 2008.
129. Denkinger MD, Weyerhauser K, Nikolaus T, Coll-Planas L. Reliability of the abbreviated version of the Late-Life Function and Disability Instrument: a meaningful and feasible tool to assess physical function and disability in the elderly. Z Gerontol Geriatr 2009;42:28–38.
130. Denkinger MD, Igl W, Coll-Planas L, Bleicher J, Nikolaus T, Jamour M. Evaluation of the short form of the late-life function and disability instrument in geriatric inpatients-validity, responsiveness, and sensitivity to change. J Am Geriatr Soc 2009;57:309–14.
131. Denkinger MD, Igl W, Jamour M, Bader A, Bailer S, Lukas A, et al. Does functional change predict the course of improvement in geriatric inpatient rehabilitation? Clin Rehabil 2010;24:463–70.
132. Hall K, McAuley E. Individual, social environmental and physical environmental barriers to achieving 10,000 steps per day among older women. Health Educ Res 2010;25:478–88.
133. Motl RW, McAuley E, Suh Y. Validity, invariance and responsiveness of a self-report measure of functional limitations and disability in multiple sclerosis. Disabil Rehabil 2010;32:1260–71.

Summary Table for Disability Measures

Scale	Purpose/ content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
Arthritis Impact Measurement Scales 2 (AIMS2)	Arthritis-specific health status	Self-administered	Moderate to high	By hand: moderate; by computer: low	High scores = poor health	Good	Good	Good	Arthritis-specific; good psychometric properties	Lengthy time to complete; scope of disability questions limited; not useful for comparative disease studies
Arthritis Impact Measurement Scales 2-Short Form (AIMS2-SF)	Arthritis-specific health status	Self-administered	Low	Low	High scores = poor health	Fair to good	Good	Good	Arthritis-specific; time to administer is short	Users should be aware of variability in subscales; scope of disability questions limited; not useful for comparative disease studies
Organization for Economic Cooperation and Development (OECD) Long-Term Disability Questionnaire (LTD)	Cross-disease disability measure aimed at measuring long-term disruptions to normal activities	Interviewer, self-administered	Low	Low	No standard scoring available	Not available	Limited	Not designed to measure change over time	Precursor to other disability measures	Has been superseded by other physical disability measures
EQ-5D	Disability-focused quality of life measure	Interviewer, self-, telephone, electronic administration	Low	Moderate to high	EQ-5D Health State is complex to create, consult EQ-5D web site	Fair to good	Fair to good	Poor to fair	Quick and easy to administer	Complex scoring; designed as broad overview measure; some validity issues
World Health Organization Disability Assessment Schedule II (WHODASII)	Disablement, including activity limitations and participation restrictions	Interviewer, self-, telephone, proxy administration	Varies with version, typically low to moderate	Moderate	Higher = more disability	Good	Good	Fair to good	Useful for comparative research; captures a range of elements important to arthritis; content goes beyond functional disability	More data needed for arthritis, especially responsiveness to change
Late-Life Function and Disability Instrument (LLFDI)	Measure of physical disablement for older adults	Interviewer, self-, telephone (computerized adaptive test) administration	Moderate	Moderate	Higher scores = better function	Fair to good	Fair to good	Fair, more data needed	Comprehensive; potential to be useful in arthritis samples	Lengthy; more data needed for arthritis; time frame (i.e., typical day) may hinder ability to detect change
Late-Life Function and Disability Instrument Abbreviated (LLFDI-Abbreviated)	Measure of physical disablement for older adults	Interviewer, self-administered	Low	Low to moderate	High scores = better function	Fair to good	Fair to good	Data limited	See Late-Life Function and Disability Instrument; quick to administer	More psychometric testing needed; timeframe (i.e., typical day) may hinder ability to detect change

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S240–S252
DOI 10.1002/acr.20543
© 2011, American College of Rheumatology

Measures of Adult Pain

Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP)

GILLIAN A. HAWKER, SAMRA MIAN, TETYANA KENDZERSKA, AND MELISSA FRENCH

INTRODUCTION

Our purpose is to provide an overview of available generic and rheumatology population–specific questionnaires suitable for evaluating pain in adult rheumatology populations. The content, ease of use, and measurement properties of the questionnaires are presented and compared in order to assist both clinicians and researchers select the questionnaire that is most appropriate for their purpose. The questionnaires are presented in the following order: generic unidimensional pain questionnaires (Visual Analog Scale and Numeric Rating Scale), generic multidimensional pain questionnaires (Short-form McGill Pain Questionnaire, Chronic Pain Grade Scale, and Short Form-36 Bodily Pain Scale), and finally an arthritis-specific pain questionnaire (Measure of Intermittent and Constant Osteoarthritis Pain). Composite measures of arthritis symptoms, including pain and associated disability, specifically the Western Ontario and McMaster Universities Osteoarthritis Index and the Arthritis Impact Measurement Scales, are described in Measures of Knee Function and Measures of Disability, respectively.

VISUAL ANALOG SCALE (VAS) FOR PAIN

Description

Purpose. The pain VAS is a unidimensional measure of pain intensity (1), which has been widely used in diverse adult populations, including those with rheumatic diseases (2–5).

Gillian A. Hawker, MD, MSc, FRCPC, Samra Mian, MSc, Tetyana Kendzerska, MD, Melissa French, MSc: University of Toronto, Toronto, Ontario, Canada.

Address correspondence to Gillian Hawker, MD, MSc, FRCPC, Canadian Osteoarthritis Research Program, 76 Grenville Street, 8th Floor, Room 815, Women's College Hospital, University of Toronto, Toronto, ON, M5S 1B2. E-mail: gillian.hawker@wchospital.ca.

Submitted for publication February 2, 2011; accepted in revised form June 20, 2011.

Content. The pain VAS is a continuous scale comprised of a horizontal (HVAS) or vertical (VVAS) line, usually 10 centimeters (100 mm) in length, anchored by 2 verbal descriptors, one for each symptom extreme (2,6). Instructions, time period for reporting, and verbal descriptor anchors have varied widely in the literature depending on intended use of the scale (7).

Number of items. The pain VAS is a single-item scale.

Response options/scale. For pain intensity, the scale is most commonly anchored by "no pain" (score of 0) and "pain as bad as it could be" or "worst imaginable pain" (score of 100 [100-mm scale]) (6–8). To avoid clustering of scores around a preferred numeric value, numbers or verbal descriptors at intermediate points are not recommended (4,9).

Recall period for items. Varies, but most commonly respondents are asked to report "current" pain intensity or pain intensity "in the last 24 hours."

Practical Application

How to obtain. The pain VAS is available in the public domain at no cost (7). Graphic formats for the VAS may be obtained from Scott & Huskisson (9) or online: http://www.amda.com/tools/library/whitepapers/hospiceinltc/appendix-a.pdf.

Method of administration. The pain VAS is self-completed by the respondent. The respondent is asked to place a line perpendicular to the VAS line at the point that represents their pain intensity (2,9,10).

Scoring. Using a ruler, the score is determined by measuring the distance (mm) on the 10-cm line between the "no pain" anchor and the patient's mark, providing a range of scores from 0–100 (6).

Score interpretation. A higher score indicates greater pain intensity. Based on the distribution of pain VAS scores in postsurgical patients (knee replacement, hysterectomy, or laparoscopic myomectomy) who described their postoperative pain intensity as none, mild, moderate, or severe, the following cut points on the pain VAS have been recommended: no pain (0–4 mm), mild pain (5–44 mm), moderate pain (45–74 mm), and severe pain (75–100 mm) (11). Normative values are not available.

Respondent burden. The VAS takes <1 minute to complete (3,7).

Administrative burden. The VAS is administered as a paper and pencil measure. As a result, it cannot be administered verbally or by phone. No training is required other than the ability to use a ruler to measure distance to determine a score (7,9). Caution is required when photocopying the scale as this may change the length of the 10-cm line (6). As slightly lower scores have been reported on the HVAS compared to the VVAS (12), the same alignment of scale should be used consistently within the same patient.

Translations/adaptations. Minimal translation difficulties have led to an unknown number of cross-cultural adaptations.

Psychometric Information

Method of development. The pain VAS originated from continuous visual analog scales developed in the field of psychology to measure well-being (13,14). Woodforde and Merskey (15) first reported use of the VAS pain scale with the descriptor extremes “no pain at all” and “my pain is as bad as it could possibly be” in patients with a variety of conditions. Subsequently, others reported use of the scale to measure pain in rheumatology patients receiving pharmacologic pain therapy (2,6,9). While variable anchor pain descriptors have been used, there does not appear to be any rationale for selecting one set of descriptors over another.

Acceptability. The pain VAS requires little training to administer and score and has been found to be acceptable to patients (2,10). However, older patients with cognitive impairment may have difficulty understanding and therefore completing the scale (6,16). Supervision during completion may minimize these errors (9).

Reliability. Test–retest reliability has been shown to be good, but higher among literate ($r = 0.94$, $P < 0.001$) than illiterate patients ($r = 0.71$, $P < 0.001$) before and after attending a rheumatology outpatient clinic (8).

Validity. In the absence of a gold standard for pain, criterion validity cannot be evaluated. For construct validity, in patients with a variety of rheumatic diseases, the pain VAS has been shown to be highly correlated with a 5-point verbal descriptive scale (“nil,” “mild,” “moderate,” “severe,” and “very severe”) and a numeric rating scale (with response options from “no pain” to “unbearable pain”), with correlations ranging from 0.71–0.78 and 0.62–0.91, respectively) (3). The correlation between vertical and horizontal orientations of the VAS is 0.99 (12).

Ability to detect change. In patients with chronic inflammatory or degenerative joint pain, the pain VAS has demonstrated sensitivity to changes in pain assessed hourly for a maximum of 4 hours and weekly for up to 4 weeks following analgesic therapy ($P < 0.001$) (10). In patients with rheumatoid arthritis, the minimal clinically significant change has been estimated as 1.1 points on an 11-point scale (or 11 points on a 100-point scale) (17). A minimum clinically important difference of 1.37 cm has been determined for a 10-cm pain VAS in patients with rotator cuff disease evaluated after 6 weeks of nonoperative treatment (18).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths/caveats and cautions/clinical and research usability. The VAS is widely used due to its simplicity and adaptability to a broad range of populations and settings. Its acceptability as a generic pain measure was demonstrated in the early 1970s. Limitations to the use of the pain VAS include the following: older patients may have difficulty completing the pain VAS due to cognitive impairments or motor skill issues, scoring is more complicated than that for the Numeric Rating Scale for pain (described below), and it cannot be administered by telephone, limiting its usefulness in research.

NUMERIC RATING SCALE (NRS) FOR PAIN

Description

Purpose. The NRS for pain is a unidimensional measure of pain intensity in adults (19–21), including those with chronic pain due to rheumatic diseases (3,8). Although various iterations exist, the most commonly used is the 11-item NRS (22), which is described here.

Content. The NRS is a segmented numeric version of the visual analog scale (VAS) in which a respondent selects a whole number (0–10 integers) that best reflects the intensity of their pain (21). The common format is a horizontal bar or line (23). Similar to the pain VAS, the NRS is anchored by terms describing pain severity extremes (3,20,21).

Number of items. The pain NRS is a single 11-point numeric scale (3).

Response options/scale. An 11-point numeric scale (NRS 11) with 0 representing one pain extreme (e.g., “no pain”) and 10 representing the other pain extreme (e.g., “pain as bad as you can imagine” and “worst pain imaginable”) (20,21).

Recall period for items. Varies, but most commonly respondents are asked to report pain intensity “in the last 24 hours” or average pain intensity (24).

Practical Application

How to obtain. Available from the web site: http://www.partnersagainstpain.com/printouts/A7012AS2.pdf.

Method of administration. The NRS can be administered verbally (therefore also by telephone) or graphically for self-completion (6). The respondent is asked to indicate the numeric value on the segmented scale that best describes their pain intensity.

Scoring. The number that the respondent indicates on the scale to rate their pain intensity is recorded. Scores range from 0–10.

Score interpretation. Higher scores indicate greater pain intensity.

Respondent burden. The pain NRS takes <1 minute to complete.

Administrative burden. The pain NRS is easy to administer and score (6,25).

Translations/adaptations. Like the pain VAS, minimal language translation difficulties support the use of the NRS across cultures and languages (26).

Psychometric Information

Method of development. To improve discrimination for detecting relatively small changes, an NRS comprised of numbers along a scale was used in a population of 100 patients with a variety of rheumatic diseases (3). Variations in pain descriptors used as anchors for end points on the pain NRS have been reported in the literature (3,6,24). However, the methodology used to develop these various anchor terms is unknown.

Acceptability. Chronic pain patients prefer the NRS over other measures of pain intensity, including the pain VAS, due to comprehensibility and ease of completion (27). However, focus groups of patients with chronic back pain and symptomatic hip and knee osteoarthritis (OA) have found that the pain NRS is inadequate in capturing the complexity and idiosyncratic nature of the pain experience or improvements due to symptom fluctuations (28,29).

Reliability. High test–retest reliability has been observed in both literate and illiterate patients with rheumatoid arthritis (r = 0.96 and 0.95, respectively) before and after medical consultation (8).

Validity. For construct validity, the NRS was shown to be highly correlated to the VAS in patients with rheumatic and other chronic pain conditions (pain >6 months): correlations range from 0.86 to 0.95 (3,8).

Ability to detect change. In clinical trials of pregabalin for diabetic neuropathy, postherpetic neuralgia, chronic low back pain, fibromyalgia, and OA, analyses of the relationships between changes in pain NRS scores and patient reports of overall improvement, measured using a standard 7-point patient global impression of change, demonstrated a reduction of 2 points, or 30%, on the pain NRS scores to be clinically important (22). Similar results were found in low back pain patients when changes in pain NRS scores were compared to patient improvements in pain after physical therapy, using a 15-point Global Rating of Change scale (19).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths/caveats and cautions/clinical and research usability. The pain NRS is a valid and reliable scale to measure pain intensity. Strengths of this measure over the pain VAS are the ability to be administered both verbally (therefore by telephone) and in writing, as well as its simplicity of scoring. However, similar to the pain VAS, the pain NRS evaluates only 1 component of the pain experience, pain intensity, and therefore does not capture the complexity and idiosyncratic nature of the pain experience or improvements due to symptom fluctuations.

MCGILL PAIN QUESTIONNAIRE (MPQ)

Description

Purpose. A multidimensional pain questionnaire designed to measure the sensory, affective and evaluative aspects of pain and pain intensity in adults with chronic pain, including pain due to rheumatic diseases (30,31).

Content. The scale contains 4 subscales evaluating the sensory, affective and evaluative, and miscellaneous aspects of pain, responses to which comprise the Pain Rating Index, and a 5-point pain intensity scale (Present Pain Intensity).

Number of items. The Pain Rating Index contains 78 pain descriptor items categorized into 20 subclasses, each containing 2–6 words that fall into 4 major subscales: sensory (subclasses 1–10), affective (subclasses 11–15), evaluative (subclass 16), and miscellaneous (subclasses 17–20). There is also a 1-item pain intensity scale (30).

Response options/scale. The value (score) associated with each descriptor is based on its position or rank order within the word set. The Present Pain Intensity scale, a measure of the magnitude of pain experienced by an individual, is a numeric-verbal combination that indicates overall pain intensity (31) and includes 6 levels: none (0), mild (1), discomforting (2), distressing (3), horrible (4), and excruciating (5) (32).

Recall period for items. Present pain (31).

Examples of use. The MPQ can be used to evaluate the efficacy and effectiveness of pain interventions and to identify qualities of pain associated with distinct nocioceptive disorders and neuropathic pain disorders, including arthritis (30).

Practical Application

How to obtain. The MPQ is available at no cost from the developer, Ronald Melzack, PhD, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, Quebec H3A 1B1, Canada, and online (http://www.qolid.org) by paying a membership fee.

Method of administration. The MPQ is interviewer-administrated using paper and pencil. The interviewer must read instructions to the respondent and define any words that the respondent does not understand. For each subclass of words, the respondent is instructed to select 1 word that fits their present pain. If none of the words describe their pain, then no word is selected (30,33).

Scoring. The MPQ is scored by hand by first counting the number of words selected to obtain a Number of Words Chosen score (0–20 words). Pain Rating Index scores range from 0–78 based on the rank values of the chosen words. The value (score) associated with each descriptor is based on its position or rank order in the word set, such that the first word is given a value of 1, the next a value of 2 and so on. Rank values are summed within each subclass as well as overall. Scores on the Present Pain Intensity scale range from 0–5 (31).

Score interpretation. A higher score on the MPQ indicates worse pain. The Pain Rating Index is interpreted both in terms of quantity of pain, as evidenced by the number of words used and the rank values of the words, as well as the quality of pain, as evidenced by the particular words that are chosen. The normative mean scores across painful conditions ranged from 24–50% of the maximum score (7).

Respondent burden. Completion of the MPQ can take up to 20 minutes (33).

Administrative burden. No training is required to score and interpret the MPQ, other than the ability of the interviewer to define each word (30,33). Time to score is 2–5 minutes.

Translations/adaptations. There are a total of 44 different versions of the MPQ, representing 26 different languages/cultures (34). The MPQ has been translated into English, French (35), German (36–38), Norwegian (39), Danish (40), Italian (41), Japanese (42), Finnish (43), Spanish (44), Chinese (45), Dutch (46), Amharic (47), Slovak (48), Turkish (49), and Portuguese (50–52).

Psychometric Information

Method of development. Pain descriptors were derived from recording the words used by chronic pain patients to describe their pain; these descriptors were then categorized into subclasses and rank ordered by intensity using a numerical scale by groups of physicians, patients, and students (31,53).

Acceptability. Some respondents have difficulty with the complexity of the vocabulary used, resulting in failure to read the instructions carefully and to see essential features (54).

Reliability. In a study of general rheumatology clinic patients, test–retest reliability for the 3 MPQ pain items (“nagging,” “aching,” and “stabbing”) ranged from a high of 0.81 for 1-day recall to a low of 0.59 for 7-day recall (55). These findings are consistent with those of other studies evaluating test–retest reliability in populations with a variety of other conditions including arthritis and other musculoskeletal conditions ($r = >0.70$) (31,56,57).

Validity. *Content validity.* Arthritis patients, regardless of their disease severity, used similar words to describe the sensory aspects of their pain. MPQ words have been shown to differentiate between 4 different circumstances of rheumatoid arthritis (RA) pain (i.e., overall pain at rest, overall pain on movement, joint pain at rest, and joint pain on movement) (58). The MPQ has the ability to detect mild pain due to the multidimensional nature of the scale and the large number of pain descriptor options (59).

Construct validity. In RA, the number of sensory and affective MPQ words selected has been positively correlated with visual analog scale (VAS) scores of severity of pain at rest and on movement ($r = 0.27$, $P < 0.01$ and $r = 0.17$, $P < 0.05$, respectively) (58). Higher Pain Rating Index scores are associated with negative affect (e.g., Minnesota Multiphasic Personality Inventory, Pain Catastrophizing Scale) (60). In knee pain or knee osteoarthritis patients, higher MPQ scores were associated with greater anxiety and depression ($r = 0.30$, $P < 0.05$ and $r = 0.31$, $P < 0.05$, respectively) (60), and greater symptoms and disability using the Western Ontario and McMaster Universities pain scale ($r = 0.34–0.38$) (61).

Ability to detect change. In clinical trials designed to evaluate the efficacy of different pain therapies on postoperative pain after general surgical and orthopedic procedures, the relative effect sizes for the MPQ-Pain Rating Index compared with a 4-point categorical verbal rating scale and a pain VAS were 1.08 (moderate) and 1.12 (good), respectively (62).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths/caveats and cautions/clinical and research usability. The MPQ is a valid and reliable tool that evaluates both the quality and quantity of pain through use of unique pain descriptors. This may be useful in epidemiologic studies and clinical trials of older patients with multimorbidity, in whom pain may arise from multiple causes. Specifically, use of the MPQ may help to identify neuropathic type pain from nocioceptive type pain. A limitation of the MPQ is the rich vocabulary required of respondents for completion. Further, sex and ethnic differences may affect selection of pain descriptors. However, the interviewer can facilitate MPQ completion by providing respondents with clear definitions of words during administration.

SHORT-FORM MCGILL PAIN QUESTIONNAIRE (SF-MPQ)

Description

Purpose. The SF-MPQ, a shorter version of the MPQ, is a multidimensional measure of perceived pain in adults with chronic pain, including pain due to rheumatic diseases (30,33).

Content. The SF-MPQ is comprised of 15 words (11 sensory and 4 affective) from the original MPQ (33).

Number of items. The Pain Rating Index is comprised of 2 subscales: 1) sensory subscale with 11 words or items and 2) affective subscale with 4 words or items, which are rated on an intensity scale as 0 = none, 1 = mild, 2 = moderate, or 3 = severe. The SF-MPQ also includes 1 item for present pain intensity and 1 item for a 10-cm visual analog scale (VAS) for average pain (33).

Recall period for items. Present time.

Examples of use. To discriminate among different pain syndromes (33,63,64) and evaluate the responsiveness of different symptoms to treatment (65,66).

Practical Application

How to obtain. See this section for the MPQ above.

Method of administration. See this section for the MPQ above.

Scoring. For the Pain Rating Index, each selected word is scored from 0 (none) to 3 (severe). The total Pain Rating Index score is obtained by summing the item scores (range 0–45). Scores on the Present Pain Intensity range from 0–5 and on the VAS from 0–10.

Score interpretation. There are no established critical cut points. As for the MPQ, a higher score indicates worse pain.

Respondent burden. The SF-MPQ takes ~2–5 minutes to complete (33).

Administrative burden. No training is required to score and interpret the SF-MPQ other than the ability of the interviewer to define each word (33). Time to score is ~1 minute.

Translations/adaptations. The SF-MPQ has been translated into the following languages: English, French, Amharic (47), Chinese (67), Czech (68), Danish (69), Farsi (70), Greek (71,72), Hebrew (73), Hindi (74), Korean (75,76), Norwegian (77), Swedish (78), Thai (79), and Turkish (49,80).

Psychometric Information

Method of development. In addition to indices of overall pain intensity (the Present Pain Intensity [31] and VAS [81]), a selection of pain descriptors representing sensory and affective categories were retained from the original version of the MPQ (31). Other than for one descriptor ("splitting"), those selected for inclusion in the SF-MPQ were those chosen by greater than one-third of patients with various types of pain (31,82–84).

Acceptability. Standardized instructions for patient completion have not been published. Some difficulties with completion have been reported and attributed to unfamiliar descriptors and unclear written instructions. However, experience in completing the SF-MPQ and verbal instructions improved completion among osteoarthritis (OA) patients (85).

Reliability. For internal consistency, using the SF-MPQ in rheumatoid arthritis (RA) and fibromyalgia patients, Cronbach's alphas were estimated at $\alpha = 0.73–0.89$ (78). In the same study (78), test–retest reliability ranged from 0.45–0.73 for 1-month and 3-month intervals. Among rheumatology patients, test–retest reliability was 0.79–0.93 at intervals of 1 to 3 days (86). In an OA population, high intraclass correlations were demonstrated for the total, sensory, affective, and average pain scores (5-day period): 0.96, 0.95, 0.88, and 0.89, respectively (85).

Validity. The SF-MPQ was found to have more content validity among patients with fibromyalgia than for those with RA. Percentage of use of 15 pain descriptors by 2 groups was significantly different for all words except "throbbing" and "punishing-cruel." The mean intensity score for each word ranged from 1.69 for "sickening" to 2.60 for "tender" in the fibromyalgia group and 1.57 for "fearful" to 2.18 for "aching" in the RA group (78). For construct validity, the SF-MPQ was found to be moderately correlated with both the Western and Ontario and McMaster Universities Osteoarthritis Index and the Short Form 36 Health Survey bodily pain scales ($r = 0.36$ and -0.36, respectively; $P < 0.01$) in 200 patients with hip and knee OA (61).

Ability to detect change. Although designed for descriptive purposes, the SF-MPQ has been found to be sensitive to the effects of pain therapies in a variety of population settings (86–88). In patients with a range of musculoskeletal conditions reporting improvements in pain after rehabilitation and surgical interventions, the Norwegian SF (NSF)-MPQ scores were found to be responsive to change (standardized response mean values >0.80): a mean improvement in NSF-MPQ total scores >5 on the 0–45 scale demonstrated a clinically important change (86). In an OA population, the minimum detectable change for total, sensory, affective, average, and current pain components have been estimated as 5.2, 4.5, 2.8, 1.4, and 1.4 cm, respectively (85).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths/caveats and cautions/clinical and research usability. The SF-MPQ is easier to use and takes less time to administer and complete than the longer form. The word choices are not as complex, and the intensity ranking of mild, moderate, and severe is better understood by patients (33). However, sufficient experience is required to adequately complete the SF-MPQ; therefore, new users require supervision during completion (85). In 2009, the short form was further revised for use in neuropathic and nonneuropathic pain conditions (SF-MPQ-2). The SF-MPQ-2 includes 7 additional symptoms relevant to neuropathic pain, for a total of 22 items with 0–10 numerical response options (89). We await further psychometric testing of this revised measure, which may play a useful role in the future with respect to identifying rheumatic disease patients with neuropathic versus nociceptive pain patterns.

CHRONIC PAIN GRADE SCALE (CPGS)

Description

Purpose. The CPGS is a multidimensional measure that assesses 2 dimensions of overall chronic pain severity: pain intensity and pain-related disability. It is suitable for use in all chronic pain conditions, including chronic musculoskeletal (MSK) and low back pain (90).

Content. Subscale scores for pain intensity and disability are combined to calculate a chronic pain grade that enables classification of chronic pain patients into 5 hierarchical categories: grades 0 (no pain) to IV (high disability-severely limiting) (90).

Number of items. The CPGS is comprised of 7 items.

Response options/scale. All items are scored on an 11-point Likert scale, with responses ranging from 0–10.

Recall period for items. Pain in the past 3–6 months.

Examples of use. The CPGS has been used in epidemiologic studies and clinical trials to evaluate and compare pain severity across groups and in response to treatment effects, and in clinical practice to improve the prognostic judgments of physicians (91–93).

Practical Application

How to obtain. Please note that the scale is available in the original reference (90), as well as directly from the author.

Method of administration. The CPGS is an interview-administered questionnaire that can also be self-completed by respondents.

Scoring. Scores are calculated for 3 subscales: the characteristic pain intensity score, which ranges from 0–100, is calculated as the mean intensity ratings for reported current, worst, and average pain; the disability score, which ranges from 0–100, is calculated as the mean rating for difficulty performing daily, social, and work activities; and the disability points score, which ranges from 0–3, is derived from a combination of ranked categories of number of disability days and disability score.

Score interpretation. The 3 subscale scores (characteristic pain intensity, disability score, and the disability points score) are used to classify subjects into 1 of the 5 pain severity grades: grade 0 for no pain, grade I for low disability-low intensity, grade II for low disability-high intensity, grade III for high disability-moderately limiting, and grade IV for high disability-severely limiting.

Respondent burden. Time to complete the CPGS does not exceed 10 minutes.

Administrative burden. The CPGS is easy to administer. Scoring is complex.

Translations/adaptations. The CPGS has been adapted into UK English (94). An Italian version has been developed to evaluate severity in chronic pain patients (95).

Psychometric Information

Method of development. Interviews were conducted with primary care patients with back pain, headache, and temporomandibular disorder (96). The development of the graded classification drew on concepts by Turk and Rudy of chronic pain severity (97,98). Two of the items used in the disability score were adapted from their Multidimensional Pain Inventory (99). The Guttman scaling method was used to develop the graded classification of chronic pain (90).

Acceptability. The CPGS is easy to understand and complete based on a high response rate (76.3%) to a postal survey sent to general practice patients in the UK (94). Among MSK chronic pain patients, missing values were only noted in <3% of each of the questions in an Italian version of the CPGS (100).

Reliability. For internal consistency, Cronbach's alpha was shown to be $\alpha = 0.74$ among patients with chronic back pain (90). In an Italian version of CPGS, Cronbach's alpha ranged from 0.81–0.89 for subscales and global scores in patients with chronic MSK pain (95). In UK general practice patients with low back pain, test–retest reliability was high after a 2-week interval (weighted $\kappa = 0.81$ [95% confidence interval 0.65–0.98]) (101).

Validity. For construct validity, cross-sectional and longitudinal studies of general practice patients have shown that higher scores on the CPGS, indicating greater chronic pain, are significantly associated with higher rates of unemployment, greater pain impact scale scores, greater use of opioid analgesics and physician visits, depressed mood, and lower self-rated health status (90,94,102). Comparisons of CPGS scores with the Short Form 36 Health Survey (SF-36) indicate that a higher chronic pain grade using the CPGS is associated with poorer physical, psychological, social, and general health as measured by the SF-36 ($P = 0.001$) (102) and worse scores on the SF-36 bodily pain scale ($\rho = -0.545$, $P < 0.0001$) (95). Spearman's correlation coefficients for the CPGS scores and the various dimensions of SF-36 were highest for the pain dimension (r = −0.71 to −0.84) and lowest for the mental health dimension (r = −0.28 to −0.38) (94).

Ability to detect change. Among patients with moderate to severe chronic MSK pain, the CPGS has been shown to be modestly responsive to changes after 12 months of treatment with an efficacious pain intervention, with standardized effect sizes for the intensity and disability subscales of 0.41 and 0.43, respectively. Among participants with chronic knee or hip pain, the standardized effect size for the CPGS intensity was 0.32 (91).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths/caveats and cautions/clinical and research usability. The CPGS is a valid and reliable tool that is useful for the evaluation of chronic MSK pain; it allows for grading of the global severity of chronic pain and therefore analysis of the qualitative changes in chronic pain over time. Further, the CPGS assesses not only aspects of the pain itself, but also the impact of the pain on daily, social, and work activities, which is a significant advantage over many other pain questionnaires. A limitation of the CPGS relative to the other scales reported here is the complexity of scoring, which renders it less useful for assessment of pain at point of care. Additionally, further research is needed to be able to compare scoring methods and cut points.

SHORT FORM 36 BODILY PAIN SCALE (SF-36 BPS)

Description

Purpose. The SF-36 BPS is 1 of 8 subscales of the Medical Outcomes Study SF-36 questionnaire (103,104), a generic measure of health status designed for use in population surveys (105). In 1996, version 2.0 of the SF-36 (SF-36v2) was introduced to correct deficiencies identified in the original version, SF-36v1 (106). The 2-item SF-36 BPS subscale assesses bodily pain as a dimension of health status (104,105).

Content. The SF-36 BPS assesses bodily pain intensity and interference of pain with normal activities.

Number of items. The SF-36 BPS is a 2-item scale.

Response options/scale. Intensity of bodily pain is evaluated using a 6-point rating scale of "none" to "very severe." The extent to which pain has interfered with work is evaluated on a 5-point rating scale from "not at all" to "extremely."

Recall period for items. The SF-36 BPS is available in both standard (4 week) and acute (1 week) recall versions (105,106).

Examples of use. The SF-36 and its subscales, including the BPS, have been used in epidemiologic studies to compare health status across populations and within population subgroups, such as in estimating the relative burden of different diseases, including rheumatic diseases (107), and differentiating the health benefits of different treatments (108).

Practical Application

How to obtain. The SF-36 and its various versions have been developed by the Rand Corporation and John E. Ware (SF-36 Health Survey, The Health Institute, New England Medical Center Hospitals, Box 345, 750 Washington Street, Boston, MA, 02111). The Medical Outcomes Trust, Health Assessment Lab, and Quality Metric Incorporated are co-copyright holders of all SF-36 and SF-12 surveys. All SF-36 survey instruments, scoring manuals, and licenses for use are available from QualityMetric at www.qualitymetric.com. Different charges are levied for academic and commercial use.

Method of administration. The SF-36 BPS is suitable for self-administration, computerized administration, or administration by a trained interviewer in person or by telephone. Telephone voice recognition interactive systems and online administrations are currently being evaluated.

Scoring. Responses for each of the 2 SF-36 BPS items are recoded into final item values (109). The raw scale score is computed as a simple algebraic sum of the recoded item values. The raw scale score is then transformed to a 0–100 scale. Norm-based scores may be calculated for SF-36v2 by including population normative data in the scoring algorithms. The BPS score is only calculated if both items are completed.

Score interpretation. SF-36v1 BPS scores range from 0–100. A higher score indicates lack of bodily pain. SF-36v2 uses norm-based scoring, where 50 is the "average" for the population. Therefore, scores above or below 50 can be considered above or below, respectively, the population average health status for bodily pain, and scores can be interpreted based on deviance from the mean (10 points = 1 SD). Population normative data are available for the US and UK.

Respondent burden. The BPS takes <2 minutes to complete.

Administrative burden. Training to administer, score, and interpret is minimal. Administration guidelines are specific and clearly outlined.

Translations/adaptations. SF-36 has been translated and adapted for use in more than 50 countries as part of the International Quality of Life Assessment (IQOLA) Project. Currently, published forms include the German (110), Spanish (111), Swedish (112), and Italian (113) translations and English-language adaptations for use in Australia/New Zealand, Canada, and the UK. Information about translations is available from the IQOLA Program of the Health Assessment Lab in Boston, Massachusetts (http://www.iqola.org).

Psychometric Information

Method of development. The Medical Outcomes Study researchers selected and adapted questionnaire items from instruments in use since the 1970s and 1980s (114–117) to develop a new 149-item Functioning and Well-Being Profile. Items were subsequently reduced and improvements were made in item wording, format, and scoring to produce the SF-36. One item on pain intensity was retained from an earlier version of the SF-20 question regarding bodily pain or discomfort. In order to improve prediction of best total scores for the Behavioural Effects of Pain Scale in the Medical Outcomes Study (114), a second item was added to measure the extent to which pain interferes in activities (105).

Acceptability. Generally easy to administer and complete (103,118).

Reliability. For internal consistency, Cronbach's α for the SF-36 BPS administered in hip (118) and knee (119) osteoarthritis (OA) patients was 0.72 and 0.77, respectively; using a Chinese version of the SF-36 BPS in rheumatoid arthritis (RA) patients, Cronbach's α was 0.91 (120). In adults with persistent back, hip, or knee pain recruited from primary care, Cronbach's α was 0.59 (100). Among patients from 2 general practices in the UK, the test–retest reliability over a 2-week period was 0.78 (121). Over a 14-day interval, test–retest reliability of a Chinese SF-36 version used in Chinese-speaking RA patients was 0.82 (120).

Validity. Regarding face and content validity, items were derived from pre-existing questionnaires used in large population studies. However, both floor and ceiling effects have been reported (118). Regarding construct validity, the proportions reporting no pain on the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) and the SF-36 BPS were 32.2% and 13.6%, respectively, and pain scores were modestly correlated (−0.55) (122) among patients who had undergone joint replacement surgery. In the same study, the WOMAC better discriminated subjects with varying severity of knee problems, whereas the SF-36 BPS better discriminated subjects with varying levels of self-reported health status and comorbidity. In patients with hip and knee OA, correlations between the WOMAC pain scale and the SF-36 BPS are in the range of 0.6–0.7 (61,121,123). In Chinese-speaking patients with RA, moderate correlations were reported between the Chinese SF-36 BPS and physician global assessment of disease activity (r = −0.34), physician's assessment of global disease activity (r = −0.35), and patient pain assessment based on a pain visual analog scale (r = −0.48) (120).

Ability to detect change. Although the SF-BPS is designed to measure the health status of populations, it has been shown to be responsive to improvements in pain. Among patients undergoing knee replacement surgery, the estimated minimum clinically important difference (MCID) ranged from 16.86/100 (SD 31.83) at 6 months to 6.69/100 (SD 29.20) at 2 years (119). In a similar study on hip replacement, the estimated MCID ranged from 14.67/100 (SD 26.46) to 18.34/100 (SD 27.06) at 6 months and 2 years, respectively (118). The minimal detectable change for the SF-36 BPS ranged from 37.91/100 (knee OA) to 38.09/100 (hip OA) at 6 months (118,119).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths/caveats and cautions/clinical and research usability. The SF-36 BPS is a valid and reliable generic questionnaire designed to evaluate bodily pain as a dimension of overall health status and has been widely used internationally and in diverse populations. Its advantages include simplicity of administration and usefulness in making comparisons across populations for research purposes. At point of patient care, a disease-specific pain measure may be more useful to discriminate levels of pain severity, and therefore response to treatment.

MEASURE OF INTERMITTENT AND CONSTANT OSTEOARTHRITIS PAIN (ICOAP)

Description

Purpose. The ICOAP measure is a multidimensional osteoarthritis (OA)-specific measure designed to comprehensively evaluate the pain experience in people with hip or knee OA, including pain intensity, frequency, and impact on mood, sleep, and quality of life, independent of the effect of pain on physical function (28). It is intended for use alongside a measure of physical disability.

Content. The ICOAP is an 11-item scale evaluating 2 pain domains: a 5-item scale evaluates constant pain and a 6-item scale evaluates intermittent pain or "pain that comes and goes" (28). Two supplementary questions can be used to assess predictability of intermittent pain when present (124). Both a hip and knee joint version of the ICOAP are available (28).

Number of items: 11 items in 2 domains with 2 supplementary items on intermittent pain predictability.

Response options/scale. All items are constructed as rating scales with 5 levels of response. For items asking about intensity, response options are "not at all," "mildly," "moderately," "severely," and "extremely." For items about frequency, response options are "never," "rarely," "sometimes," "often," and "very often" (28). For the supplementary items asking about predictability of pain, the response options are "never," "rarely," "sometimes," "often," and "very often" (124).

Recall period for items. Past week.

Endorsements. Osteoarthritis Research Society International (OARSI)/Outcome Measures in Rheumatology (OMERACT) Initiative.

Practical Application

How to obtain. The ICOAP and ICOAP Users Guide can be obtained free of charge from the OARSI web site, www.oarsi.org.

Method of administration. The measure can be interviewer-administered in person or by telephone (28). Respondents should complete both subscales (63).

Scoring. Each ICOAP item is scored from 0–4. A score is separately produced for each subscale by summing the items' subscale scores and then normalizing each score from 0 (no pain) to 100 (extreme pain). A total ICOAP score can be calculated by summing the 2 subscale scores, and normalizing from 0 (no pain) to 100 (extreme pain). Rules have been created to deal with missing data (63). No scoring guidelines are available for the 2 supplementary items.

Score interpretation. Higher scores indicate a worse pain experience.

Respondent burden. The ICOAP takes <10 minutes to complete.

Administrative burden. Easy to administer and score.

Translations/adaptations. To date, the ICOAP has been translated into the following languages: English (North America and UK), Czech, Dutch, French (France), German (125), Italian, Norwegian, Spanish (Castillan), North and Central American Spanish, Swedish, Portuguese (50), Greek, Romanian, and Russian (126). Translated versions are available at www.oarsi.org.

Psychometric Information

Method of development. Focus groups were conducted in individuals with painful hip or knee OA in 4 countries (US, UK, Canada, and Australia) to generate items pertaining to the OA pain experience (28,127). Content analysis of resulting transcripts was used to identify themes, which were verified with participants. Subsequent psychometric testing was conducted in subjects age ≥40 years with hip or knee OA drawn from rheumatologists' practices, joint replacement wait lists, and from among the members of an existing OA cohort (28).

Acceptability. The ICOAP has been shown to be easy to understand and complete; subjects felt positive about the inclusion of the 2 distinct pain domains (constant pain and pain that comes and goes) (126).

Reliability. Regarding content validity, Cronbach's α was 0.93 (28) for 100 individuals with hip and knee OA. Test–retest reliability in 76 individuals with hip and knee OA, age ≥40 years, demonstrated an intraclass correlation coefficient of 0.85 (95% confidence interval 0.76–0.91) (122).

Validity. Content and face validity were determined through focus groups used to develop the ICOAP. For construct validity, descriptive analyses of items demonstrated good distribution of response options across all items (28). Total and subscale ICOAP scores are significantly correlated with scores on the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain scale, the Knee Injury and OA Outcome Score (KOOS) symptoms scale, and self-rated effect of hip/knee problems on quality of life, with Spearman's correlation coefficients ranging in magnitude from 0.60 (KOOS symptoms) to 0.81 (WOMAC pain scale) (28).

Ability to detect change. The ICOAP has been found to be responsive to changes in OA pain in response to pharmacologic interventions (128) and joint replacement surgery (129). For the knee, standard response means (SRMs) ranged from 0.49–0.57 for the ICOAP intermittent, constant, and total scores comparable to that for the WOMAC pain (SRM 0.54). For the hip, SRMs ranged from 0.11–0.19, again comparable to that for the WOMAC (SRM 0.15) (128). The ICOAP detected large improvements in pain resulting from joint replacement surgery with SRMs (0.84–1.02 for knee replacement and 1.50–2.29 for hip replacement) (129).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths/caveats and cautions/clinical and research usability. The ICOAP is a valid and reliable measure that is unique in that it is intended to evaluate the multidimensional pain experience in OA, distinct from the impact of pain on physical functioning. Prior experience with the WOMAC, the most commonly used OA measure, has shown high correlations between the pain and physical function subscales. As a result, evaluation of OA pain using the WOMAC is confounded by physical disability. The ICOAP is intended for use together with a measure of OA disability. Although the ICOAP has been translated from English into a number of languages, only a few have been evaluated for validity, reliability, and responsiveness.

DISCUSSION

There are multiple measures available to assess pain in adult rheumatology populations. Each measure has its own strengths and weaknesses. Both the Visual Analog Scale for Pain and the Numeric Rating Scale (NRS) for Pain are unidimensional single-item scales that provide an estimate of patients' pain intensity. They are easy to administer, complete, and score. Of the 2, the pain NRS may be preferred at point of patient care due to simpler scoring. In research, the pain NRS may similarly be preferred due to its ability to be administered both verbally and in writing. However, neither measure provides a comprehensive evaluation of pain in patients with rheumatic disease. To evaluate the multiple dimensions of acute and chronic pain, a number of valid and reliable questionnaires are available. The McGill Pain Questionnaire (MPQ) is a generic pain measure useful largely for research purposes to describe not only the quantity (intensity), but also the quality of the patients' pain. The Chronic Pain Grade Scale (CPGS) is similarly a generic pain measure useful for research purposes to describe, evaluate and compare chronic pain severity (its intensity and impact) across groups and in response to treatment effects. The third generic multidimensional pain measure, the Short Form-36 Bodily Pain Scale (SF-36 BPS), is useful in evaluating pain in the context of overall health status, and therefore most suitable for use in making comparisons across populations and between subgroups within populations. Unlike the MPQ and CPGS, the SF-36 BPS is simple enough for use at point of care. Finally, the Measure of Intermittent and Constant Osteoarthritis Pain is an osteoarthritis-specific pain measure that is recommended for descriptive and evaluative purposes in both clinical practice and research to provide a comprehensive evaluation of the pain experience in osteoarthritis, including the impact of pain on mood, sleep and quality of life, separate and distinct from the impact of pain on functioning. Due to the variability in purpose, content, method of administration, respondent and administrative burden, and evidence to support the psychometric properties of each measure, no one pain measure can be recommended for use in all situations. We encourage clinicians and researchers to use this information presented in this chapter to help guide the selection of the questionnaire that is most appropriate for their specific purpose.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. McCormack HM, Horne DJ, Sheather S. Clinical applications of visual analogue scales: a critical review. Psychol Med 1988;18:1007–19.
2. Huskisson EC. Measurement of pain. Lancet 1974;2:1127–31.
3. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. Ann Rheum Dis 1978;37:378–81.
4. Huskisson EC, Wojtulewski JA, Berry H, Scott J, Hart FD, Balme HW. Treatment of rheumatoid arthritis with fenoprofen: comparison with aspirin. Br Med J 1974;1:176–80.
5. Berry H, Huskisson EC. Treatment of rheumatoid arthritis. Clin Trials J 1972;4:13–5.
6. Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. Pain 1986;27:117–26.
7. Burckhardt CS, Jones KD. Adult measures of pain: The McGill Pain Questionnaire (MPQ), Rheumatoid Arthritis Pain Scale (RAPS), Short-Form McGill Pain Questionnaire (SF-MPQ), Verbal Descriptive Scale (VDS), Visual Analog Scale (VAS), and West Haven-Yale Multidisciplinary Pain Inventory (WHYMPI). Arthritis Rheum 2003;49:S96–104.
8. Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 1990;17:1022–4.
9. Scott J, Huskisson EC. Graphic representation of pain. Pain 1976;2:175–84.
10. Joyce CR, Zutshi DW, Hrubes VF, Mason RM. Comparison of fixed interval and visual analogue scales for rating chronic pain. Eur J Clin Pharmacol 1975;8:415–20.
11. Jensen MP, Chen C, Brugger AM. Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. J Pain 2003;4:407–14.
12. Scott J, Huskisson EC. Vertical or horizontal visual analogue scales. Ann Rheum Dis 1979;38:560.
13. Aitken RC. Measurement of feelings using visual analogue scales. Proc R Soc Med 1969;62:989–93.
14. Clarke PR, Spear FG. Reliability and sensitivity in the self-assessment of well-being. [abstract]. Bull Br Psychol Soc 1964;17:18A.
15. Woodforde JM, Merskey H. Some relationships between subjective measures of pain. J Psychosom Res 1972;16:173–8.
16. Kremer E, Atkinson JH, Ignelzi RJ. Measurement of pain: patient preference does not confound pain measurement. Pain 1981;10:241–8.
17. Wolfe F, Michaud K. Assessment of pain in rheumatoid arthritis: minimal clinically significant difference, predictors, and the effect of anti-tumor necrosis factor therapy. J Rheumatol 2007;34:1674–83.
18. Tashjian RZ, Deloach J, Porucznik CA, Powell AP. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. J Shoulder Elbow Surg 2009;18:927–32.
19. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. Spine 2005;30:1331–4.
20. Jensen MP, McFarland CA. Increasing the reliability and validity of pain intensity measurement in chronic pain patients. Pain 1993;55:195–203.
21. Rodriguez CS. Pain measurement in the elderly: a review. Pain Manag Nurs 2001;2:38–46.
22. Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001;94:149–58.

23. Johnson C. Measuring pain. Visual analog scale versus numeric pain scale: what is the difference? J Chiropr Med 2005;4:43–4.
24. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9–19.
25. Jensen MP, Karoly P, O'Riordan EF, Bland F Jr, Burns RS. The subjective experience of acute pain. Clin J Pain 1989;5:153–9.
26. Langley GB, Sheppeard H. The visual analogue scale: its use in pain measurement. Rheumatol Int 1985;5:145–8.
27. De C Williams AC, Davies HT, Chadury Y. Simple pain rating scales hide complex idiosyncratic meanings. Pain 2000;85:457–63.
28. Hawker GA, Davis AM, French MR, Cibere J, Jordan JM, March L, et al. Development and preliminary psychometric testing of a new OA pain measure: an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:409–14.
29. Hush JM, Refshauge KM, Sullivan G, De Souza L, McAuley JH. Do numerical rating scales and the Roland-Morris Disability Questionnaire capture changes that are meaningful to patients with persistent back pain? Clin Rehabil 2010;24:648–57.
30. Burckhardt CS. The use of the McGill Pain Questionnaire in assessing arthritis pain. Pain 1984;19:305–14.
31. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. Pain 1975;1:277–99.
32. Escalante A, Lichtenstein MJ, White K, Rios N, Hazuda HP. A method for scoring the pain map of the McGill pain questionnaire for use in epidemiologic studies. Aging Clin Exp Res 1995;7:358–66.
33. Melzack R. The short-form McGill Pain Questionnaire. Pain 1987;30:191–7.
34. Menezes CL, Maher CG, McAuley JH, Costa LO. Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. J Clin Epidemiol 2009;62:934–43.
35. Boureau F, Luu M, Doubrere JF, Gay C. Construction of a questionnaire for the self-evaluation of pain using a list of qualifiers. Therapie 1984;39:119–29.
36. Geissner E. Measuring pain via questionnaires: several results concerning the validity of a modified German version of the McGill Pain Questionnaire. Z Klin Psychol 1988;17:334–40.
37. Kiss I, Muller H, Abel M. The McGill Pain Questionnaire: German version. Pain 1987;29:195–207.
38. Stein C, Mendl G. The German counterpart to McGill Pain Questionnaire. Pain 1988;32:251–5.
39. Kim HS, Schwartz-Barcott D, Holter IM, Lorensen M. Developing a translation of the McGill pain questionnaire for cross-cultural comparison: an example from Norway. J Adv Nurs 1995;21:421–6.
40. Van Lankveld WF, van Pad Bosch PF, van de Putte LF, van der Staak CF, Naring G. Pain in rheumatoid arthritis measured with the visual analogue scale and the Dutch version of the McGill Pain Questionnaire. Ned Tijdschr Geneeskd 1992;136:1166–70.
41. Maiani G, Sanavio E. Semantics of pain in Italy: the Italian version of the McGill Pain Questionnaire. Pain 1985;22:399–405.
42. Hasegawa M, Hattori S, Mishima M, Matsumoto IF, Kimura T, Baba Y, et al. The McGill Pain Questionnaire, Japanese version, reconsidered: confirming the theoretical structure. Pain Res Manag 2001;6:173–80.
43. Ketovuori H, Pontinen PJ. A pain vocabulary in Finnish: the Finnish pain questionnaire. Pain 1981;11:247–53.
44. Lazaro C, Bosch F, Torrubia R, Banos JE. The development of a Spanish questionnaire for assessing pain: preliminary data concerning reliability and validity. Euro J Psychol Assess 1994;10:145–51.
45. Hui YL, Chen AC. Analysis of headache in a Chinese patient population. Ma Zui Xue Za Zhi 1989;27:13–8.
46. Van der Kloot WA, Oostendorp RA, van der Meij J, van den HJ. The Dutch version of the McGill pain questionnaire: a reliable pain questionnaire. Ned Tijdschr Geneeskd 1995;139:669–73. In Dutch.
47. Aboud FE, Hiwot MG, Arega A, Molla M, Samson SN, Seyoum N, et al. The McGill Pain Questionnaire in Amharic: Zwai Health Center patients' reports on the experience of pain. Ethiop Med J 2003;41:45–61.
48. Bartko D, Kondas M, Janco S. Quantification of pain in neurology. Cesk Neurol Neurochir 1984;47:113–21.
49. Kuguoglu S, Aslan FE, Olgun N. Turkish version of the McGill Melzack Pain Questionnaire Form (MPQF). Agri-Istanbul 2003;15:47–51.
50. Goncalves RS, Cabri J, Pinheiro JP, Ferreira PL, Gil J. Cross-cultural adaptation and validation of the Portuguese version of the intermittent and constant osteoarthritis pain (ICOAP) measure for the knee. Osteoarthritis Cartilage 2010;18:1058–61.
51. Pimenta CA, Teixeiro MJ. Proposal to adapt the McGill Pain Questionnaire into Portuguese. Rev Esc Enferm USP 1996;30:473–83. In Portuguese.
52. Varoli FK, Pedrazzi V. Adapted version of the McGill Pain Questionnaire to Brazilian Portuguese. Braz Dent J 2006;17:328–35.
53. Torgerson WS. Theory and methods of scaling. New York: Wiley and Son; 1958.
54. Chapman CR, Casey KL, Dubner R, Foley KM, Gracely RH, Reading AE. Pain measurement: an overview. Pain 1985;22:1–31.
55. Broderick JE, Schwartz JE, Vikingstad G, Pribbernow M, Grossman S, Stone AA. The accuracy of pain and fatigue items across different reporting periods. Pain 2008;139:146–57.
56. Love A, Leboeuf C, Crisp TC. Chiropractic chronic low back pain sufferers and self-report assessment methods. Part I. A reliability study of the Visual Analogue Scale, the Pain Drawing and the McGill Pain Questionnaire. J Manipulative Physiol Ther 1989;12:21–5.
57. Roche PA, Klestov AC, Heim HM. Description of stable pain in rheumatoid arthritis: a 6 year study. J Rheumatol 2003;30:1733–8.
58. Papageorgiou AC, Badley EM. The quality of pain in arthritis: the words patients use to describe overall pain and pain in individual joints at rest and on movement. J Rheumatol 1989;16:106–12.
59. Katz J, Clairoux M, Kavanagh BP, Roger S, Nierenberg H, Redahan C, et al. Pre-emptive lumbar epidural anesthesia reduces postoperative pain and patient-controlled morphine consumption after lower abdominal surgery. Pain 1994;59:395–403.
60. Creamer P, Lethbridge-Cejku M, Hochberg MC. Determinants of pain severity in knee osteoarthritis: effect of demographic and psychosocial variables using 3 pain measures. J Rheumatol 1999;26:1785–92.
61. Gandhi R, Tsvetkov D, Dhottar H, Davey JR, Mahomed NN. Quantifying the pain experience in hip and knee osteoarthritis. Pain Res Manag 2010;15:224–8.
62. Jenkinson C, Carroll D, Egerton M, Frankland T, McQuay H, Nagle C. Comparison of the sensitivity to change of long and short form pain measures. Qual Life Res 1995;4:353–7.
63. ICOAP User's Guide. Osteoarthritis Research Society International. 2011. URL: http://www.oarsi.org/pdfs/pain_indexes/ICOAP_USERS_GUIDE_07072010.pdf.
64. Rasmussen PV, Sindrup SH, Jensen TS, Bach FW. Symptoms and signs in patients with suspected neuropathic pain. Pain 2004;110:461–9.
65. Dworkin RH, Corbin AE, Young JP Jr, Sharma U, LaMoreaux L, Bockbrader H, et al. Pregabalin for the treatment of postherpetic neuralgia: a randomized, placebo-controlled trial. Neurology 2003;60:1274–83.
66. Gilron I, Bailey JM, Tu D, Holden RR, Weaver DF, Houlden RL. Morphine, gabapentin, or their combination for neuropathic pain. N Engl J Med 2005;352:1324–34.
67. Hsieh LL, Kuo CH, Yen MF, Chen TH. A randomized controlled clinical trial for low back pain treated by acupressure and physical therapy. Prev Med 2004;39:168–76.
68. Knotek P, Solcova I, Zalsky M. Czech version of the short form McGill Pain Questionnaire: restandardization. Bolest 2002;5:169–72.
69. Perkins FM, Werner MU, Persson F, Holte K, Jensen TS, Kehlet H. Development and validation of a brief, descriptive Danish pain questionnaire (BDDPQ). Acta Anaesthesiol Scand 2004;48:486–90.
70. Najafi-Ghezeljeh T, Ekman I, Nikravesh MY, Emami A. Adaptation and validation of the Iranian version of Angina Pectoris characteristics questionnaire. Int J Nurs Pract 2008;14:470–6.
71. Georgoudis G, Watson PJ, Oldham JA. The development and validation of a Greek version of the short-form McGill Pain Questionnaire. Eur J Pain 2000;4:275–81.
72. Georgoudis G, Oldham JA, Watson PJ. Reliability and sensitivity measures of the Greek version of the short form of the McGill Pain Questionnaire. Eur J Pain 2001;5:109–18.
73. Sloman R, Rosen G, Rom M, Shir Y. Nurses' assessment of pain in surgical patients. J Adv Nurs 2005;52:125–32.
74. Ahuja S, Saluja V, Bhattacharya A. A Modified Short form McGill Pain Questionnaire for evaluation of post-operative pain and behavioural response to pain relief. J Anaesthesiol Clin Pharmacol 1999;15:149–53.
75. Lee HF, Nicholson LL, Adams RD, Maher CG, Halaki M, Bae SS. Development and psychometric testing of Korean language versions of 4 neck pain and disability questionnaires. Spine (Phila Pa 1976) 1976;31:1841–5.
76. Lee MC, Essoka G. Patient's perception of pain: comparison between Korean-American and Euro-American obstetric patients. J Cult Divers 1998;5:29–37.
77. Ljunggren AE, Strand LI, Johnsen TB. Development of the Norwegian short-form McGill Pain Questionnaire (NSF-MPQ). Adv Physiother 2007;9:169–80.
78. Burckhardt CS, Bjelle A. A Swedish version of the short-form McGill Pain Questionnaire. Scand J Rheumatol 1994;23:77–81.
79. Kitisomprayoonkul W, Klaphajone J, Kovindha A. Thai Short-form McGill Pain Questionnaire. J Med Assoc Thai 2006;89:846–53.
80. Yakut Y, Yakut E, Bayar K, Uygur F. Reliability and validity of the Turkish version short-form McGill pain questionnaire in patients with rheumatoid arthritis. Clin Rheumatol 2007;26:1083–7.
81. Huskisson EC. Current practice in rheumatology. Practitioner 1983;227:1087.
82. Grushka M, Sessle BJ. Applicability of the McGill Pain Questionnaire to the differentiation of 'toothache' pain. Pain 1984;19:49–57.
83. Hunter M. The Headache Scale: a new approach to the assessment of headache pain based on pain descriptions. Pain 1983;16:361–73.

84. Leavitt JW. “Science” enters the birthing room: obstetrics in America since the eighteenth century. J Am Hist 1983;70:281–304.
85. Grafton KV, Foster NE, Wright CC. Test-retest reliability of the Short-Form McGill Pain Questionnaire: assessment of intraclass correlation coefficients and limits of agreement in patients with osteoarthritis. Clin J Pain 2005;21:73–82.
86. Strand LI, Ljunggren AE, Bogen B, Ask T, Johnsen TB. The Short-Form McGill Pain Questionnaire as an outcome measure: test-retest reliability and responsiveness to change. Eur J Pain 2008;12:917–25.
87. Birch S, Jamison RN. Controlled trial of Japanese acupuncture for chronic myofascial neck pain: assessment of specific and nonspecific effects of treatment. Clin J Pain 1998;14:248–55.
88. Ruoff GE, Rosenthal N, Jordan D, Karim R, Kamin M. Tramadol/acetaminophen combination tablets for the treatment of chronic lower back pain: a multicenter, randomized, double-blind, placebo-controlled outpatient study. Clin Ther 2003;25:1123–41.
89. Dworkin RH, Turk DC, Revicki DA, Coyne KS, Peirce-Sandner S, Burke LB, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). Pain 2009;144:35–42.
90. Von Korff M, Ormel J, Keefe FJ, Dworkin SF. Grading the severity of chronic pain. Pain 1992;50:133–49.
91. Elliott AM, Smith BH, Smith WC, Chambers WA. Changes in chronic pain severity over time: the Chronic Pain Grade as a valid measure. Pain 2000;88:303–8.
92. Elliott AM, Smith BH, Penny KI, Smith WC, Chambers WA. The epidemiology of chronic pain in the community. Lancet 1999;354:1248–52.
93. Von Korff M, Stewart WF, Lipton RB. Assessing headache severity. Neurology 1994;44:S40–6.
94. Smith BH, Penny KI, Purves AM, Munro C, Wilson B, Grimshaw J, et al. The Chronic Pain Grade questionnaire: validation and reliability in postal research. Pain 1997;71:141–7.
95. Salaffi FF, Stancati AF, Grassi W. Reliability and validity of the Italian version of the Chronic Pain Grade questionnaire in patients with musculoskeletal disorders. Clin Rheumatol 2006;25:619–31.
96. Von Korff M, Dworkin SF, Le RL. Graded chronic pain status: an epidemiologic evaluation. Pain 1990;40:279–91.
97. Turk DC, Rudy TE. Towards a comprehensive assessment of chronic pain patients. Behav Res Ther 1987;25:237–49.
98. Turk DC, Rudy TE. Toward an empirically derived taxonomy of chronic pain patients: integration of psychological assessment data. J Consult Clin Psychol 1988;56:233–8.
99. Kerns RD, Turk DC, Rudy TE. The West Haven-Yale Multidimensional Pain Inventory (WHYMPI). Pain 1985;23:345–56.
100. Krebs EE, Bair MJ, Damush TM, Tu W, Wu J, Kroenke K. Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. Med Care 2010;48:1007–14.
101. Dunn KM, Jordan K, Croft PR. Does questionnaire structure influence response in postal surveys? J Clin Epidemiol 2003;56:10–6.
102. Penny KI, Purves AM, Smith BH, Chambers WA, Smith WC. Relationship between the chronic pain grade and measures of physical, social and psychological well-being. Pain 1999;79:275–9.
103. Ware JE, Kosinski M, Keller SK. SF-36 physical and mental health summary scales: a user’s manual. Boston: The Health Institute; 1994.
104. McHorney CA, Ware JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36) II: psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247–63.
105. Ware JE Jr, Sherbourne CD. The MOS 36-item Short-Form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
106. Ware JE, Kosinski M, Dewey JE. How to score version two of the SF-36 Health Survey. Lincoln (RI): QualityMetric; 2000.
107. Ware JE, Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) project. J Clin Epidemiol 1998;51:903–12.
108. Shiely JC, Bayliss MS, Keller SD, Tsai C, Ware JE. SF-36 Health Survey annotated bibliography: the first edition (1988-1995). Boston: The Health Institute, New England Medical Center; 1996.
109. Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey: manual and interpretation guide. Boston: The Health Institute, New England Medical Center; 1993.
110. Bullinger M. German translation and psychometric testing of the SF-36 Health Survey: preliminary results from the IQOLA Project. Soc Sci Med 1995;41:1359–66.
111. Alonso J, Prieto L, Anto JM. The Spanish version of the SF-36 Health Survey (the SF-36 health questionnaire): an instrument for measuring clinical results. Med Clin (Barc) 1995;104:771–6. In Spanish.
112. Sullivan M, Karlsson J, Ware JE Jr. SF-36 Halsoenkat: Svensk manual och tolkningsguide (Swedish manual and interpretation guide). Gothenburg: Sahlgrenska University Hospital; 1994.
113. Apolone G, Cifani S, Liberati MC, Mosconi P. Questionario sullo stato di salute SF-36. Traduzione e validazione della versione italiana: risultati del progetto IQOLA. Metodologia e Didattica Clinica 1997;5:86–94.
114. Stewart AL, Ware JE. Measuring functioning and well-being: the Medical Outcomes Study approach. Durham (NC): Duke University Press; 1992.
115. Dupuy HJ. The Psychological General Well-Being (PGWB) Index. In: Wenger NK, Mattson ME, Furberg JF, Elinson JA, editors. Assessment of quality of life in clinical trials of cardiovascular therapies. New York: Le Jacq; 1984. p. 170–83.
116. Hulka BS, Cassel JC. The AAFP-UNC study of the organization, utilization, and assessment of primary medical care. Am J Public Health 1973;63:494–501.
117. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. Health Serv Res 1973;8:228–45.
118. Quintana JM, Escobar A, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. Osteoarthritis Cartilage 2005;13:1076–83.
119. Escobar A, Quintana JM, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. Osteoarthritis Cartilage 2007;15:273–80.
120. Koh ET, Leong KP, Tsou IY, Lim VH, Pong LY, Chong SY, et al. The reliability, validity and sensitivity to change of the Chinese version of SF-36 in oriental patients with rheumatoid arthritis. Rheumatology (Oxford) 2006;45:1023–8.
121. Brazier JE, Harper R, Jones NM, O’Cathain A, Thomas KJ, Usherwood T, et al. Validating the SF-36 Health Survey questionnaire: new outcome measure for primary care. BMJ 1992;305:160–4.
122. Bombardier C, Melfi C, Paul J, Green R, Hawker GA, Wright J, et al. Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. Med Care 1995;33:AS131–44.
123. Salaffi F, Leardini G, Canesi B, Mannoni A, Fioravanti A, Caporali R, et al. Reliability and validity of the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index in Italian patients with osteoarthritis of the knee. Osteoarthritis Cartilage 2003;11:551–60.
124. Hawker GA, French MR, Elkayam JG, Davis AM. Unpredictability of intermittent knee OA pain: impact on pain, function, and mood [abstract]. Arthritis Rheum 2010;62 Suppl 10:S284–5.
125. Kessler S, Grammozis A, Gunther KP, Kirschner S. The Intermittent and Constant Pain Score (ICOAP): a questionnaire to assess pain in patients with gonarthritis. Z Orthop Unfall 2010;149:22–6. In German.
126. Maillefert JF, Kloppenburg M, Fernandes L, Punzi L, Gunther KP, Martin Mola E, et al. Multi-language translation and cross-cultural adaptation of the OARSI/OMERACT measure of intermittent and constant osteoarthritis pain (ICOAP). Osteoarthritis Cartilage 2009;17:1293–6.
127. Hawker GA, Stewart L, French MR, Cibere J, Jordan JM, March L, et al. Understanding the pain experience in hip and knee osteoarthritis: an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:415–22.
128. Hawker GA, Davis A, Lohmander S. Responsiveness of the new OARSI-OMERACT OA pain and function measures [abstract]. Ann Rheum Dis 2009;68 Suppl 3:350.
129. Davis AM, Lohmander LS, Wong R, Venkataramanan V, Hawker GA. Evaluating the responsiveness of the ICOAP following hip or knee replacement. Osteoarthritis Cartilage 2010;18:1043–5.

Summary Table for Adult Pain Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
VAS	Unidimensional measure of pain intensity One horizontal or vertical line with varying time points and descriptor anchors	Self-administered Pencil/paper	<1 minute	No training required Ability to use a ruler to measure distance to score VAS required	Higher scores indicate greater pain intensity	Good; Test–retest reliability higher among literate than illiterate rheumatology outpatients (r = 0.94 and r = 0.71, respectively)	Good; VAS scores shown to be highly correlated with other pain measure scores Range r = 0.62–0.91	Excellent; Sensitive to measuring changes in pain associated with treatment or time. (MCID of 1.37 cm on 10-cm pain VAS; change of 1.1 points on 11-point scale)	Acceptable/valid/ reliable measure of pain intensity	May not be appropriate for use in older and/ or illiterate population
NRS	Unidimensional measure of pain intensity A single 11-point scale that consists of a numeric segmented horizontal bar or line	Self-administered or interviewer-administered (by telephone)	<1 minute	No training required to administer and score	Higher scores indicate greater pain intensity No established cut points	Excellent; Test–retest reliability high for both literate and illiterate rheumatology outpatients (r = 0.96 and r = 0.95 respectively)	Excellent; Construct validity: NRS highly correlated to VAS in patients with rheumatic and other chronic pain conditions. (r = >0.86)	Excellent; A reduction of 2 points, or 30%, on NRS scores is clinically important	Acceptable/valid/ reliable measure of pain intensity Can be administered by telephone	Does not capture the complex and idiosyncratic nature of pain experience or improvements due to symptom fluctuations
MPQ	A multidimensional measure of sensory, affective and evaluative aspects of pain and pain intensity Comprised of 78 pain descriptor items and one-item pain intensity scale	Interviewer-administered Pencil/paper	<20 minutes	No training required to score (2–5 minutes to score) Interviewer must have ability to describe each pain descriptor	Higher scores on MPQ indicate worse pain No established cut points	Good; Test–retest reliability in rheumatology populations is adequate (r = >0.70)	Good; Construct validity: MPQ words positively correlated with pain VAS scores at rest and on the movement in rheumatoid arthritis patients (range 0.17–0.27). In knee OA patients, higher MPQ scores associated with greater anxiety, depression, symptoms, and disability using the WOMAC (r = >0.30)	Good; ES for MPQ-Pain Rating Index compared with 4-point categorical verbal rating scale and a pain VAS were 1.08 (moderate) to 1.12 (good)	Acceptable/valid/ reliable measure of quality and quantity of pain using unique pain descriptors	Rich vocabulary of respondents required for completion Ethnic and sex differences may affect selection of pain descriptors
SF-MPQ	A multidimensional measure of perceived pain intensity Comprised of 15 pain descriptor items	Interviewer-administered Pencil/paper	2–5 minutes	No training required to score (1 minute to score) Interviewer must have ability to describe each pain descriptor	Higher scores on SF-MPQ indicate worse pain No established cut points	Good; Internal consistency: Cronbach's alpha in rheumatology population ranged from 0.73 to 0.89. Test–retest reliability ranged from 0.45 to 0.73 at 1- and 3-month intervals. For 1–3 day intervals, test–retest was high (range: 0.79–0.93). ICC was high for each SF-MPQ subscale, VAS average pain and total scores (ICC >0.89)	Good; Content validity: The mean intensity scores for each SF-MPQ words ranged from 1.57 to 2.60 in rheumatology patients. In OA patients, SF-MPQ is moderately correlated with both WOMAC and SF-26 bodily pain scales (r = 0.36 and r = −0.36, respectively)	Good; In MSK patients, SRM values for Norwegian SF-MPQ were >0.80. Clinically important change: mean improvement in total scores >5 on the 0–45 NSF-MPQ scale Minimum detectable change for total, sensory, affective, average and current pain are 5.2, 4.5, 2.8, 1.4, 1.4 cm	Acceptable/valid/ reliable short-form version of the MPQ Easier to use, more understandable, and takes less time to administer and complete than the longer form	Supervision during completion required for new users
CPGS	A multidimensional pain measure that assesses 2 dimensions of overall chronic pain severity: pain intensity and pain-related disability Subscale scores for pain intensity and disability are combined to calculate a chronic pain grade and classify chronic pain patients into 5 hierarchical categories: grades 0 (no pain) to IV (high disability-severely limiting)	Self-administered or interviewer-administered	<10 minutes	No training required Easy to administer but scoring is complex	3 subscale scores are used to classify subjects into 1 of 5 pain severity categories	Good; Internal consistency: Cronbach's α = 0.74 in low back pain patients. In chronic MSK pain patients, Cronbach's α ranged from 0.81–0.89 for an Italian CPGS Test–retest reliability was high after 2-week interval in UK general practice patients (weighted κ = 0.81 95% CI 0.65–0.98)	Good; Construct validity: Spearman's correlation coefficients for CPGS and SF-36 were high for the pain dimension (r = −0.71 to −0.84) and low for the mental health dimension (r = −0.28 to −0.38)	Good; Effect sizes for CPGS intensity and disability subscales in chronic MSK patients were 0.41 and 0.43, respectively. In chronic knee or hip pain, effect size was 0.32	Acceptable, valid and reliable measure of chronic MSK pain. CPGS allows for the grading of global severity of chronic pain and qualitative changes in chronic pain over time	Complexity of scoring limits its use for assessment of pain at point of care

(continued)

Summary Table *(Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
SF-36 BPS	A 2-item bodily pain scale that measures pain intensity and pain interference with normal activities. The SF-36 BP is 1 of the 8 SF-36 subscales	Self-administered, computer-administered, or interviewer-administered (in-person or by telephone)	<2 minutes	Minimal training required to administer, and score. (User's Guide available)	Higher scores indicate lack of bodily pain	Excellent; Internal consistency: Cronbach's α in hip and knee OA patients was 0.72 and 0.77, respectively. Using a Chinese version, Cronbach's α in RA patients was 0.91. Test–retest reliability was 0.78 in general practice patients and 0.82 in Chinese-speaking RA patients over a 2-week period	Good; Construct validity: proportion of patients reporting no pain on WOMAC and BPS were 32.2% and 13.6%. Correlations between WOMAC pain scale and SF-36 BPS ranged from 0.6–0.7	Acceptable; MCID in patients undergoing knee replacement surgery ranged from 16.86/100 (SD 31.83) at 6 months to 6.69/100 (SD 29.20) at 2 years. For hip replacement, MCID ranged from 14.67/100 (SD 26.46) to 18.34/100 (SD 27.06) at 6 months and 2 years respectively. MDC ranged from 37.91/100 to 38.09/100 at 6 months for hip and knee, respectively	Acceptable, valid and reliable generic measure of bodily pain that is simple to administer and use in diverse populations	May not be a useful measure to discriminate levels of pain severity and thus response to treatment
ICOAP	Multidimensional scale that comprehensively evaluates the pain experience in people with hip or knee OA, independent of the effect of pain on physical function An 11-item scale evaluating 2 pain domains: constant pain and intermittent pain (pain that comes and goes) Two supplementary questions can be used to assess the predictability of intermittent pain, when present	Interviewer-administered (in person or by telephone) Respondents should complete both subscales	<10 minutes	No training required to administer and score. (User's Guide available)	Higher scores indicate worse pain experience	Excellent; Internal consistency: Cronbach's α was 0.93 in hip and knee OA subjects. Test–retest reliability was high in hip and knee OA subjects (ICC 0.85)	Good; Construct validity: Spearman's correlation coefficients for ICOAP scores ranged from 0.60 to 0.81 for WOMAC and KOOS, respectively	Good; For knee OA, SRMs ranged from 0.49–0.57 for ICOAP subscales. For hip OA, SRMs ranged from 0.11– 0.19. SRMs ranged from 0.84–1.02 for knee replacement and 1.50–2.29 for hip replacement	Acceptable/valid/reliable measure of the multi-dimensional pain experience in OA, distinct from the impact of pain on physical functioning	Only a few translated versions of the ICOAP have been assessed for validity, reliability, and responsiveness

* VAS = visual analog scale; MCID = minimum clinically important difference; NRS = numeric rating scale; MPQ = McGill Pain Questionnaire; OA = osteoarthritis; WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index; ES = effect size; SF-MPQ = Short Form MPQ; ICC = intraclass correlation coefficient; MSK = musculoskeletal; SRM = standardized response mean; NSF = Norwegian Short Form; CPGS = Chronic Pain Grade Scale; 95% CI = 95% confidence interval; SF-36 BPS = Short Form 36 Bodily Pain Scale; RA = rheumatoid arthritis; MDC = minimal detectable change; ICOAP = Measure of Intermittent and Constant Osteoarthritis Pain; KOOS = Knee Injury and Osteoarthritis Outcome Score.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S446–S453
DOI 10.1002/acr.20559

Measures of Health-Related Quality of Life in Pediatric Systemic Lupus Erythematosus

Childhood Health Assessment Questionnaire (C-HAQ), Child Health Questionnaire (CHQ), Pediatric Quality of Life Inventory Generic Core Module (PedsQL-GC), Pediatric Quality of Life Inventory Rheumatology Module (PedsQL-RM), and Simple Measure of Impact of Lupus Erythematosus in Youngsters (SMILEY)

AIMEE HERSH

INTRODUCTION

Measures of physical function/disability and health-related quality of life (HRQOL) have become critical determinants of outcomes for patients with pediatric-onset systemic lupus erythematosus (SLE) (1–3). HRQOL has been defined as a "multi-dimensional concept that includes the physical, psychological and social functioning associated with an illness or its treatment" (4). At least one measure of HRQOL has been included in recent studies aimed to develop a core set of variables to assess flare criteria and response to therapy in pediatric SLE (5–7). The Childhood Health Assessment Questionnaire, which is a measure of physical function/disability, and several HRQOL instruments (Child Health Questionnaire, Pediatric Quality of Life Inventory Generic Core Module, Pediatric Quality of Life Inventory Rheumatology Module, and Simple Measure of Impact of Lupus Erythematosus in Youngsters) have been validated in pediatric SLE, and will be discussed here. This review includes both a description of the measures' content as well as their psychometric properties as it relates to pediatric SLE.

CHILDHOOD HEALTH ASSESSMENT QUESTIONNAIRE (C-HAQ)

Description

Purpose. The C-HAQ measures functional health status for pediatric patients 6 months to 18 years of age with a chronic rheumatic disease. The original publication in 1994 described validation of the C-HAQ for patients with juvenile idiopathic arthritis (JIA) (8). Revised versions of the C-HAQ have been suggested for JIA (9,10); this review pertains to the original version of the C-HAQ.

Content. The C-HAQ assesses disability in 8 domains, including dressing and grooming, arising, eating, walking, hygiene, reach, grip, and activities. The C-HAQ also includes visual analog scale (VAS) scores for pain and overall well-being.

Number of items. In the disability index, there are 30 items in 8 categories. Each category has 2–5 component items. In addition, questions are asked about assistive devices or personal aids needed to perform the 30 functions. The pain and overall well-being scales are individual questions.

Response options/scale. For each category in the disability index, the respondent selects the amount of difficulty the child may have with a particular task as a result of their condition. Each item is rated from 0–3 with 0 = "without any difficulty," 1 = "with some difficulty," 2 = "with much difficulty," and 3 = "unable to do." If the item is not applicable to the subject (i.e., they would not be expected to perform that particular task due to young age), then the parent would select "Not applicable." The highest score for any component question within a category determines the score for the category. Use of assistive devices or a personal aide automatically increases the score to a 2. If a component score is left blank, the score for that category is determined by the remaining completed questions. If all component questions are left blank, that category is left blank. Pain severity is measured on a VAS with 0 = no pain and 100 = very severe pain, and a score from 0–3 is determined based on the location of the respondent's mark. Well-being ("rate how your child is doing") is measured on a VAS scale with 0 = very well and 100 = very poor.

Recall period for items. Questions pertain to the week preceding the assessment.

Dr. Hersh's work was supported by the American College of Rheumatology Research and Education Foundation Rheumatology Investigator Award.

Aimee Hersh, MD: University of Utah, Salt Lake City.

Address correspondence to Aimee Hersh, MD, Department of Pediatrics, Division of Allergy, Immunology & Rheumatology, University of Utah, 295 Chipeta Way, Salt Lake City, UT 84108. E-mail: aimee.hersh@hsc.utah.edu.

Submitted for publication March 8, 2011; accepted in revised form July 8, 2011.

Practical Application

How to obtain. A copy of the C-HAQ can be obtained at the following URL: http://aramis.stanford.edu/index.html.

Method of administration. Interview (in person or telephone) or self-administered by either the child or the parent. Moderate correlations between child and parent reports have been demonstrated in pediatric systemic lupus erythematosus (SLE) (11).

Scoring. There are specific scoring instructions for the C-HAQ.

Score interpretation. The C-HAQ score is calculated as the mean of the 8 category scores in the disability index. Scores range from 0–3; higher scores reflect more disability. There are no normative values for the C-HAQ in healthy children. The minimum clinically important differences (MCIDs) for the C-HAQ in patients with JIA ranged between a score improvement of −0.188 and a score worsening of +0.125 (12).

Respondent burden. The C-HAQ takes <10 minutes to complete.

Administrative burden. The C-HAQ takes <10 minutes to administer and <5 minutes to score.

Translations/adaptations. The C-HAQ has been cross-culturally adapted and validated in multiple languages.

Psychometric Information

Method of development. The C-HAQ was adapted from the adult Stanford Health Assessment Questionnaire (8). There is at least 1 question per domain relevant to children of all ages. The face validity of the instrument was evaluated by a group of 20 health professionals and the parents of 22 healthy children.

Acceptability, reliability, and validity in pediatric SLE. A cross-sectional study of 24 pediatric SLE patients assessed the correlation between C-HAQ scores and disease activity utilizing the SLE Disease Activity Index (SLEDAI) and Physician's Global Assessment (PGA). The mean ± SD child-report C-HAQ score was 0.35 ± 0.35 (median 0.3), and the mean ± SD parent-report score was 0.14 ± 0.2 (median 0) (11). The median SLEDAI score for the cohort was 4; both the SLEDAI and C-HAQ exhibited a floor effect. The C-HAQ correlated moderately with the SLEDAI ($\rho = 0.4$, $P = 0.04$) but did not correlate with PGA. The C-HAQ was also validated in a study of 504 patients with active pediatric SLE who were assessed at baseline (prior to major therapeutic intervention) and then at 6-month followup (13). Mean parent-report C-HAQ scores were 0.83 ± 0.94 at baseline and 0.19 ± 0.43 at followup. In this cohort, subjects had a high degree of disease activity as evidenced by a mean ± SD SLEDAI score of 18.12 ± 10.14 at baseline; SLEDAI scores improved to 6.21 ± 6.46 at followup. Several measures of disease activity, including laboratory parameters and HRQOL, were collected at the 2 time points. In evaluating reliability, the C-HAQ demonstrated excellent internal consistency in this cohort (Cronbach's $\alpha = 0.96$). With regard to construct validity, there was a moderate correlation for the absolute change from baseline to 6 months (Spearman's correlation coefficient 0.4–0.7) between the C-HAQ and the Child Health Questionnaire (CHQ) physical health score, the parent's global assessment of pain and overall well-being, and the Systemic Lupus Activity Measure. There was poor correlation (Spearman's correlation coefficient of <0.4) with PGA, European Consensus Lupus Activity Measure, 24-hour proteinuria, SLEDAI, and CHQ psychosocial health scores.

Ability to detect change. In this study of 504 patients with active SLE, the standardized response mean for the C-HAQ was moderate at 0.74. The C-HAQ demonstrated a significant ability ($P < 0.0001$) to discriminate between subjects who were improved and those who were not improved at 6 months based on the Paediatric Rheumatology International Trials Organisation/American College of Rheumatology juvenile SLE definition of improvement.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The C-HAQ is easy to administer and is the most widely used measure of health status, physical function, and disability for patients with a chronic rheumatic disease. With regard to pediatric SLE, the C-HAQ has excellent responsiveness over time, particularly among patients with more active SLE.

Caveats and cautions. The C-HAQ focuses on physical function, and may fail to capture other physical symptoms and dimensions of SLE activity that affect health status (e.g., fatigue). In addition, the C-HAQ has been criticized for its focus on "disability" versus "ability."

Clinical usability. The C-HAQ has a significant floor effect, particularly among patients with less active SLE, limiting its clinical utility. In addition, the MCID for pediatric SLE has not been established.

Research usability. The C-HAQ has demonstrated a reasonable ability to discriminate between patients who have and have not achieved clinical improvement, making it a useful tool in a research setting.

CHILD HEALTH QUESTIONNAIRE (CHQ)

Description

Purpose. Modeled after the adult Short Form 36, the CHQ is a generic measure of pediatric health-related quality of life (HRQOL) designed to measure the physical, emotional, and social components of health. The original manuscript describing the development of the CHQ was published in 1998 (14).

Content. The CHQ comprises 14 domains, including physical functioning, bodily pain or discomfort, general health, change in health, limitations in schoolwork and activities with friends, mental health, behavior, self-esteem, family cohesion, limitations in family activities, and emotional or time impact on the parent (15).

Number of items. The parent form is available in a 50- (PF-50) or 28-item (PF-28) version. The child self-report version (CHQ-CF87) consists of 87 items and is for children ≥10 years of age. From the PF-50, physical health (PhS) and psychosocial health (PsS) summary scores can be derived.

Response options/scale. Each item is scored on a Likert-type scale with higher scores indicating better or more positive health states.

Recall period for items. The change in health subscale is "compared to last year" and no recall period is used for the general health and family cohesion subscales, otherwise the subscales refer to the preceding 4-week period.

Practical Application

How to obtain. A copy of the CHQ can be obtained at the following URL: http://www.healthact.com/survey-chq.php.

Method of administration. Self-administered by the child or a parent.

Scoring. The CHQ can be hand scored or a computer scoring program can be used.

Score interpretation. Scores for each subscale range from 0–100, with higher scores reflecting better health status. These scores are standardized with a mean ± SD of 50 ± 10. Higher scores indicate higher HRQOL. Normative values for the different versions of the CHQ, and for different populations (e.g., US, Italy, Mexico), are published (16–18). The mean ± SD national norms for the CHQ PhS is 53.00 ± 8.8 and the PsS is 51.20 ± 9.1 for a population sample of US children. Although the minimal clinically important difference for the CHQ PhS has not been established in pediatric systemic lupus erythematosus (SLE), a single study reported a mean ± SD decrease in CHQ PhS scores among patients who had a disease flare (n = 89 episodes) of −2.35 ± 1.14, while episodes not associated with a flare (n = 438) had an increase in CHQ PhS scores of 0.78 ± 0.51 ($P = 0.013$) (7).

Respondent burden. The estimated time to completion depends on the length of the survey (CHQ PF-50, 10–15 minutes; PF-28, 5–10 minutes; and the CHQ-CF87, 16–25 minutes).

Adminstrative burden. Not reported.

Translations/adaptions. The CHQ has been validated in multiple languages/countries. A full list of the available translations is available at URL: http://www.healthact.com/translation-chq.php.

Psychometric Information

Method of development. The CHQ was developed as a part of the Child Assessment Project, an effort that was initiated in 1990 to develop methods for "measuring the physical and psychosocial health status and well-being of children and adolescents" (14).

Acceptablity, reliability, and validity in pediatric SLE. In a cross-sectional and multinational study, Ruperto and colleagues assessed the HRQOL of 297 pediatric SLE patients utilizing the CHQ (15). For this cohort, the mean ± SD summary PhS score was 40.2 ± 15 and the mean ± SD summary PsS score was 44.8 ± 10.7. These scores were similar to previously reported CHQ scores for patients with juvenile idiopathic arthritis, but significantly below the mean scores for the healthy control populations used in the study. In this cohort, the SLE Disease Activity Index (SLEDAI) score was significantly correlated with the CHQ PhS score (r = −0.29, $P < 0.0001$) and the CHQ PsS score (r = −0.25, $P < 0.0001$). The Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index (SDI) correlated with the CHQ PhS score (r = −0.23, $P = 0.0001$) but not the CHQ PsS score. In a study examining the relationship between HRQOL and disease course in pediatric SLE, 98 patients were followed every 3 months for up to 18 months (549 total visits) (19). The mean ± SD summary CHQ PhS score was 41.8 ± 12 and the mean ± SD summary CHQ PsS score was 49.2 ± 6.6, which was significantly below the mean scores for the normative population of healthy children. At baseline, the mean ± SD SLEDAI score was 4.8 ± 4.4 and mean ± SD SDI score was 0.42 ± 0.1. To assess the relationship between HRQOL and disease activity, subjects were grouped into 1 of 3 groups of disease activity based on British Isles Lupus Assessment Group (BILAG) score. Subjects with a BILAG score ≤1 had "inactive" or "minimally active" disease, subjects with a BILAG score >1 but ≤5 were classified as "somewhat active" disease, and subjects with a BILAG score >5 had "very active" disease. Higher disease activity (SLEDAI or BILAG score) was associated with lower CHQ PhS and CHQ PsS scores; however, there was no significant difference in CHQ PhS scores between the groups with "somewhat" and "very active disease," and the CHQ PsS failed to differentiate between the groups of patients with different levels of disease activity. Patients with minimal or absent disease damage (SDI ≤1) had significantly higher CHQ PhS, but not CHQ PsS scores.

Ability to detect change. In the 3-month interval between study visits, as compared to physician-rated worsening or improvement of disease, the scores of the CHQ PhS changed significantly ($P < 0.0005$), but the CHQ PsS did not. The standardized response mean (SRM) was <0.4 for both measures. The CHQ PhS SRM improved to 0.57 when correlated to patient-/parent-related worsening of health. Brunner et al included the CHQ PhS as the primary measure of HRQOL in a study designed to validate criteria for the evaluation of response to therapy in pediatric SLE (6). This study included 98 children who were evaluated every 3 months for up to 7 visits (623 total visits). The CHQ PhS was one of the 5 SLE core response variables obtained at each visit. The mean ± SD score at baseline was 42.4 ± 12.14. There was no significant change in the CHQ PhS among the patients who were classified as "improved" during the study time period. In a related study designed to develop flare criteria for pediatric SLE, a combination of physician-rated disease activity, a validated disease activity index (e.g., SLEDAI, BILAG), and change in CHQ PhS (not weighted) were found to be adequate to identify SLE flares in the cohort (area under the curve: 0.81, sensitivity 64%, specificity 86%) (7).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CHQ is a comprehensive measure for assessing HRQOL. The PhS and PsS summary scores are useful tools for simplifying the multiple domains measured by the CHQ. In pediatric SLE, the CHQ PhS correlates well with cross-sectional measures of both disease activity and damage, while the CHQ PsS only correlates with disease activity.

Caveats and cautions. The CHQ PsS does not discriminate well between patients with different levels of disease activity, and is not responsive to worsening or improvement of disease over time. The CHQ PhS appears to be a more accurate measure for discriminating between patients who have had an increase in disease activity (i.e., disease flare) versus those who have a decrease in disease activity.

Clinical usability. Given its relatively low respondent burden and the response of the CHQ PhS to change in disease activity over time, the CHQ may be a useful tool for measuring HRQOL in a clinical setting.

Research usability. The CHQ PhS summary score appears to be a useful tool in identifying increased disease activity (i.e., flares) among study subjects with pediatric SLE, so it may be more useful as a research tool in observational studies versus therapeutic trials, where response to therapy is the primary outcome.

PEDIATRIC QUALITY OF LIFE INVENTORY GENERIC CORE MODULE (PEDSQL-GC)

Description

Purpose. The PedsQL-GC is a generic pediatric measure of health-related quality of life (HRQOL).

Content. The questionnaire encompasses 4 health domains: physical functioning, emotional functioning, social functioning, and school functioning.

Number of items. The questionnaire contains 23 items. Physical and psychosocial health summary scores can be calculated.

Response options/scale. Items are scored using a 5-point Likert scale (0 = never a problem, 1= almost never a problem, 2 = sometimes a problem, 3 = often a problem, and 4 = almost always a problem).

Recall period for items. Questions refer to the preceding 4 weeks.

Examples of use. Disease-specific modules are available for rheumatology, asthma, diabetes mellitus, cancer, and cardiac conditions.

Practical Application

How to obtain. A copy of the PedsQL-GC can be obtained at the following web site: http://www.pedsql.org/contact.html.

Method of administration. Includes parallel child/adolescent self-report (ages 5–18) or parent proxy report (ages 2–18).

Scoring. There are specific scoring instructions.

Score interpretation. From the sum of the raw scores from the 23 items, a summary score ranging from 0–100 can be calculated, with higher scores indicating higher HRQOL. Mean ± SD normative values for a population of school-age US children were 80.64 ± 13.34 for the child self-report total score, and 76.92 ± 16.81 for the parent proxy-report total score (20). The minimum clinically important difference for the child self-report score in a diverse pediatric population was a change of 4.4, and for the parent proxy report was a change of 4.5 (21).

Respondent burden. The survey takes <4 minutes to complete.

Administrative burden. The time to administer the survey is <10 minutes. The PedsQL-GC is described as "easy to score," but no specific time period is recorded.

Translation/adaptations. The PedsQL-GC has been translated into multiple languages, available at URL: http://www.pedsql.org/translations.html.

Psychometric Information

Method of development. The current PedsQL-GC 4.0 represents the fourth version of the PedsQL-GC tool. It has been field tested with children and adolescents in multiple settings.

Acceptability, reliability, and validity in pediatric systemic lupus erythematosus (SLE). In a cross-sectional study of 24 patients with pediatric SLE, the mean ± SD summary score for the PedsQL-GC parent proxy report was 69 ± 18 and the mean ± SD child self-report score was 67 ± 20 (11). No significant correlation was found between the PedsQL-GC and measures of disease activity, including the SLE Disease Activity Index (SLEDAI) or Physician's Global Assessment (PGA). Mild correlation was found between the PedsQL-GC and the the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index (SDI). The corresponding child-parent pair (n = 19) responses were significant for the PedsQL-GC ($\rho = 0.7$, $P = 0.001$); the intraclass correlation was 0.7 (confidence interval 0.4–0.8). In a cross-sectional study of 98 pediatric SLE patients designed to assess HRQOL and its relationship to disease activity, the mean ± SD summary score for the PedsQL-GC parent proxy report was 74.6 ± 16.7 and the mean ± SD child self-report score was 78.1 ± 15; both scores were significantly lower than the mean for the normative population (US population sample of healthy children) (19). Higher disease activity, as measured by the British Isles Lupus Assessment Group (BILAG), was associated with lower PedsQL-GC scores ($P < 0.05$). With regard to disease damage, patients with minimal or no disease damage (SDI score ≤1) had higher PedsQL-GC scores than patients with more than minimal disease damage (SDI score >1; $P < 0.05$).

Ability to detect change. With assessment of change in disease status over time by either physician-rated worsening/improvement or parent-/patient-rated change in health, there was no significant change in PedsQL-GC parent proxy-report scores and patient self-report scores. The standardized response mean was <0.4, indicating a moderate response.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PedsQL-GC is a relatively brief and easily administered generic tool for measuring HRQOL.

Caveats and cautions. The PedsQL-GC scores correlate with BILAG and SDI scores, but not with SLEDAI scores or

PGA, and it may limit the ability to assess change in disease status over time.

Clinical usability. The PedsQL-GC may be useful in the clinical setting because of the low respondent burden; however, its usefulness in measuring HRQOL over time is unclear.

Research usability. Given its limited ability to assess change in disease activity over time, the PedsQL-GC summary score may have more limited utility in the research setting.

PEDIATRIC QUALITY OF LIFE INVENTORY RHEUMATOLOGY MODULE (PEDSQL-RM)

Description

Purpose. The PedsQL-RM is used in combination with the Pediatric Quality of Life Inventory Generic Core Module (PedsQL-GC), and is a pediatric rheumatology–specific measure of health-related quality of life (HRQOL).

Content. The PedsQL-RM is a brief parallel patient- and parent-report questionnaire designed for children ages 2–18 that encompasses 5 domains: pain and hurt, daily activities, treatment, worry, and communication. Parent report of the toddler age group (2–4 years) does not include a worry and communication domain.

Number of items. The PedsQL-RM 3.0 is a 22-item questionnaire.

Response options/scale. Items are scored using a 5-point Likert scale (never, almost never, sometimes, often, and always).

Recall period for items. Questions refer to the preceding 4 weeks.

Practical Application

How to obtain. A copy of the PedsQL-RM can be obtained at the following web site: http://www.pedsql.org/contact.html.

Method of administration. Child self-report (ages 5–18) or parent proxy-report (ages 2–18) questionnaire.

Scoring. Items are scored on a 5-point Likert scale (never, almost never, sometimes, often, and always).

Score interpretation. From the sum of the raw scores from the 22 items, a summary score ranging from 0–100 can be calculated, with higher scores indicating higher HRQOL. Normative values are available for patients with juvenile idiopathic arthritis (JIA) (22).

Respondent burden. Time to complete is <4 minutes.

Administrative burden. Time to administer is <10 minutes. Time to score is not reported.

Translations/adaptations. The PedsQL-RM has been translated into multiple languages, available at URL: http://www.pedsql.org/translations.html.

Psychometric Information

Method of development. PedsQL-RM was designed to measure pediatric rheumatology–specific HRQOL. The original study designed to demonstrate the reliability, validity, and responsiveness of the PedsQL-RM included 231 children and 244 parents recruited from a pediatric rheumatology clinic (23).

Acceptability, realiability, and validity in pediatric systemic lupus erythematosus (SLE). In a cross-sectional study of 24 patients with pediatric SLE, PedsQL-RM summary scores were not reported but the mean ± SD child self-report and parent proxy-report means for each of the subscales were as follows: daily activities, 92 ± 13 and 95 ± 9; treatment, 84 ± 13 and 76 ± 19; pain and hurt, 66 ± 22 and 68 ± 24; communication, 63 ± 30 and 65 ± 40; and worry, 56 ± 34 and 52 ± 28, respectively (11). The only significant correlation with disease activity was between the parent-report worry subscale and the SLE Disease Activity Index (SLEDAI; $\rho = 0.53$, $P = 0.02$). No correlation was found with disease damage. Correlations between the corresponding child-parent pair (n = 19) responses were significant for the worry ($\rho = 0.5$, $P = 0.05$) and pain and hurt domains ($\rho = 0.55$, $P = 0.02$). The intraclass correlation was 0.5 (confidence interval 0.04–0.7). In a cross-sectional study of 98 pediatric SLE patients designed to assess HRQOL and its relationship to disease activity, the mean ± SD summary score for the PedsQL-RM parent proxy report was 79.4 ± 14.3 and the mean ± SD child self-report score was 80.8 ± 14.1; although both scores were lower than the mean for the normative population (children with JIA), the differences were not statistically significant (19). To assess the relationship between HRQOL and disease activity, subjects were grouped into 1 of 3 groups of disease activity based on British Isles Lupus Assessment Group (BILAG) scores. Higher disease activity was associated with lower PedsQL-RM scores, although the difference in scores between the somewhat active and very active disease groups for the PedsQL-RM child self-report did not reach significance. With regard to disease damage, patients with minimal or no disease damage (Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index [SDI] score ≤1) had higher PedsQL-RM scores than patients with more than minimal disease damage (SDI score >1).

Ability to detect change. There was a significant change with the PedsQL-RM parent proxy-report scores and child self-report scores regardless of the method used to assess change in health status (physician-rated worsening/improvement or parent-/patient-related change in health). The standardized response mean was <0.4, indicating a moderate response.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PedsQL-RM appears to have both concurrent and construct validity, and is more responsive than the PedsQL-GC to clinically important changes in pediatric SLE.

Caveats and cautions. The PedsQL-RM summary scores correlate with BILAG and SDI scores, but not with SLEDAI scores or physician's global assessment. The PedsQL-RM does not distinguish between the highest levels of disease activity on the BILAG.

Clinical usability. The PedsQL-RM is quick and easy to complete, making it a potentially useful measure in the clinical setting.

Research usability. The Peds QL-RM has favorable psychometric properties in pediatric SLE and may be a useful tool to assess response to therapy in observational and clinical trials.

SIMPLE MEASURE OF IMPACT OF LUPUS ERYTHEMATOSUS IN YOUNGSTERS (SMILEY)

Description

Purpose. The SMILEY is the only disease-specific measure of health-related quality of life (HRQOL) in pediatric systemic lupus erythematosus (SLE).

Content. The survey captures 4 domains: effect on self (5 items), limitations (8 items), social (4 items), and burden of SLE (7 items).

Number of items. The SMILEY is a 26-item survey. The first 2 survey items are summary questions that are not included in the final score. Item 1 relates to current HRQOL status and item 2 relates to current SLE status.

Response options/scale. Responses are in the form of a pictorial 5-step scale with different facial expressions.

Recall period for items. Responses apply to the previous month.

Practical Application

How to obtain. Contact the developer: L. Nandini Moorthy, MD, MS, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey.

Method of administration. The SMILEY is a self-administered questionnaire for children with SLE <19 years of age, and is completed by both the parent and the child.

Scoring. There are specific scoring instructions. If more than 12 questions are not answered, SMILEY cannot be scored.

Score interpretation. All items, including the first 2 summary questions, score from 1–5. The total score is transformed to a 1–100 scale. Higher scores reflect higher quality of life. Because this is a disease-specific tool, there are no normative values for the SMILEY.

Respondent burden. SMILEY is at the fifth-grade reading level and takes <10 minutes to complete.

Administrative burden. SMILEY takes 10 minutes to administer and ≤10 minutes to score.

Translations/adaptations. The SMILEY has been translated into multiple languages (24).

Psychometric Information

Method of development. The 5-step scale used in SMILEY was modified with permission from the Wong-Baker FACES Pain Rating Scale. Children with SLE and their parents were involved in the different states of development of SMILEY (25).

Acceptability, reliability, and validity in pediatric SLE. SMILEY was validated in a cohort of 86 pediatric SLE patients (26). In this cohort, the mean ± SD child-report score was 65 ± 13 (range 37–93) and the mean ± SD parent-report score was 62 ± 16 (range 28–98). The median SLE Disease Activity Index (SLEDAI) was 4 (range 0–23) and the median Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index (SDI) was 1 (range 0–10). The interrater reliability for total score via intraclass correlation was 0.7–0.9 for both the parent and child total scores. Cronbach's alpha for internal consistency was 0.9 for both parent and child total scores. For the SMILEY child and SMILEY parent, correlation with other HRQOL measures were as follows: Childhood Health Assessment Questionnaire, r = 0.6 and r = 0.5; global quality of life (QOL), r = 0.5 and r = 0.6; and Pediatric Quality of Life Inventory Generic Core Module (PedsQL-GC), r = 0.6 and r = 0.6, respectively. For the Pediatric Quality of Life Inventory Rheumatology Module, the correlation for the SMILEY child report and SMILEY parent report by domain were: pain and hurt, r = 0.5 and r = 0.5; daily activities, r = 0.4 and r = 0.4; treatment, r = 0.5 and r = 0.6; worry, r = 0.6 and r = 0.5; and communication, r = 0.5 and r = 0.5, respectively. All *P* values were ≤0.001. Significant Spearman's correlations (r ≥ 0.4) were seen between child and parent total SMILEY scores and items 1–3 and 5–6 of the impact scale of the PedsQL-GC family information form, the self-concept scale, self-perceived global QOL, and self-perceived global SLE status. The child SMILEY limitation domain had mild correlation (r = 0.3) with physician's global assessment (PGA), SLEDAI (r = 0.2), and SDI (r = 0.2). There was no significant correlation with the total child and parent SMILEY scores and SLEDAI, PGA, SDI, or disease duration. SMILEY total and domain scores were higher in subjects with lower SLEDAI scores, lower PGA scores, lower SDI scores, and in those who had never used immunomodulatory therapy, including cyclophosphamide.

Ability to detect change. In a longitudinal study, 68 pediatric SLE patients were assessed at baseline and 52 patients (76%) were assessed at followup (27). There were no significant difference in SLEDAI or SDI scores between the 2 time points. With regard to the child-report SMILEY, changes in the total scores correlated with changes in patient/parent assessment of global HRQOL (r = 0.3, *P* = 0.02), patient/parent assessment of SLE status (r = 0.4, *P* = 0.002), SLEDAI (r = −0.3, *P* = 0.01), and SDI (r = −0.4, *P* = 0.005). Changes in SLEDAI and SDI corresponded most strongly with changes in the "burden of SLE" domain. Changes in the parent-report SMILEY scores correlated with changes in the patient/parent assessment of global HRQOL (r = 0.3, *P* = 0.02), patient/parent assessment of SLE status (r= 0.4, *P* = 0.002), and SDI (r = −0.3, *P* = 0.05). Changes with SDI correlated with the limitation domain. Changes in parent-report SMILEY total and domain scores did not correlate with changes in PGA and SLEDAI scores.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. SMILEY is the only HRQOL measurement tool designed specifically for patients with pediatric SLE. The use of the FACES scale may make it more accessible for use with younger patients.

Caveats and cautions. The total parent and child SMILEY scores exhibit mild/moderate correlation with markers of disease activity and damage, potentially limiting its ability to predict a change in disease activity. Additional studies are needed to evaluate the performance of the SMILEY over time.

Clinical usability. The SMILEY is reasonably short and straightforward, making it feasible for use in clinical practice.

Research usability. The SMILEY has very good psychometric properties; the burden of SLE and limitation domains may be particularly useful in measuring response to therapy in clinical trials.

DISCUSSION

Measures of physical function and health-related quality of life (HRQOL) should be included in the assessment of short- and long-term outcomes of systemic lupus erythematosus (SLE). Although several measures exist, they vary in their correlation with and their ability to predict change in disease activity over time. The Childhood Health Assessment Questionnaire is the primary measure used to assess physical function and disability in pediatric SLE, but its utility is limited due to its floor effect and its emphasis on SLE symptoms related to arthritis. Multiple measures have been used to assess HRQOL in pediatric SLE; it has been proposed that the Child Health Questionnaire physical health summary score be included in criteria for measuring global SLE flare. SMILEY, which is the only disease-specific measure of HRQOL, appears to be an effective tool for measuring multidimensional HRQOL in pediatric SLE.

AUTHOR CONTRIBUTIONS

Dr. Hersh drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Brunner HI, Giannini EH. Health-related quality of life in children with rheumatic diseases. Curr Opin Rheumatol 2003;15:602–12.
2. Moorthy LN, Peterson M, Onel KB, Harrison MJ, Lehman TJ. Quality of life in children with systemic lupus erythematosus. Curr Rheumatol Rep 2005;7:447–52.
3. Ravelli A, Ruperto N, Martini A. Outcome in juvenile onset systemic lupus erythematosus. Curr Opin Rheumatol 2005;17:568–73.
4. Miller ML, LeBovidge J, Feldman B. Health-related quality of life in children with arthritis. Rheum Dis Clin North Am 2002;28:493–501, vi.
5. Ruperto N, Ravelli A, Cuttica R, Espada G, Ozen S, Porras O, et al, for the Pediatric Rheumatology International Trials Organization (PRINTO) and the Pediatric Rheumatology Collaborative Study Group (PRCSG). The Pediatric Rheumatology International Trials Organization criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: prospective validation of the disease activity core set. Arthritis Rheum 2005;52:2854–64.
6. Brunner HI, Higgins GC, Wiers K, Lapidus SK, Olson JC, Onel K, et al. Prospective validation of the provisional criteria for the evaluation of response to therapy in childhood-onset systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2010;62:335–44.
7. Brunner HI, Klein-Gitelman MS, Higgins GC, Lapidus SK, Levy DM, Eberhard A, et al. Toward the development of criteria for global flares in juvenile systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2010;62:811–20.
8. Singh G, Athreya BH, Fries JF, Goldsmith DP. Measurement of health status in children with juvenile rheumatoid arthritis. Arthritis Rheum 1994;37:1761–9.
9. Groen W, Unal E, Norgaard M, Maillard S, Scott J, Berggren K, et al. Comparing different revisions of the Childhood Health Assessment Questionnaire to reduce the ceiling effect and improve score distribution: data from a multi-center European cohort study of children with JIA. Pediatr Rheumatol Online J 2010;8:16.
10. Lam C, Young N, Marwaha J, McLimont M, Feldman BM. Revised versions of the Childhood Health Assessment Questionnaire (C-HAQ) are more sensitive and suffer less from a ceiling effect. Arthritis Rheum 2004;51:881–9.
11. Moorthy LN, Harrison MJ, Peterson M, Onel KB, Lehman TJ. Relationship of quality of life and physical function measures with disease activity in children with systemic lupus erythematosus. Lupus 2005; 14:280–7.
12. Brunner HI, Klein-Gitelman MS, Miller MJ, Barron A, Baldwin N, Trombley M, et al. Minimal clinically important differences of the childhood health assessment questionnaire. J Rheumatol 2005;32: 150–61.
13. Meiorin S, Pistorio A, Ravelli A, Iusan SM, Filocamo G, Trail L, et al. Validation of the Childhood Health Assessment Questionnaire in active juvenile systemic lupus erythematosus. Arthritis Rheum 2008;59: 1112–9.
14. Landgraf JM, Maunsell E, Speechley KN, Bullinger M, Campbell S, Abetz L, et al. Canadian-French, German and UK versions of the Child Health Questionnaire: methodology and preliminary item scaling results. Qual Life Res 1998;7:433–45.
15. Ruperto N, Buratti S, Duarte-Salazar C, Pistorio A, Reiff A, Bernstein B, et al. Health-related quality of life in juvenile-onset systemic lupus erythematosus and its relationship to disease activity and damage. Arthritis Rheum 2004;51:458–64.
16. Landgraf JM, Abetz L, Ware JE. The CHQ user's manual. 1st ed. Boston: The Health Institute, New England Medical Center; 1996.
17. Duarte C, Ruperto N, Goycochea MV, Maldonado R, Beristain R, De Inocencio J, et al. The Mexican version of the Childhood Health Assessment Questionnaire (C-HAQ) and the Child Health Questionnaire (CHQ). Clin Exp Rheumatol 2001;19 Suppl 23:S106–10.
18. Ruperto N, Ravelli A, Pistorio A, Malattia C, Viola S, Cavuto S, et al. The Italian version of the Childhood Health Assessment Questionnaire (C-HAQ) and the Child Health Questionnaire (CHQ). Clin Exp Rheumatol 2001;19 Suppl 23:S91–5.
19. Brunner HI, Higgins GC, Wiers K, Lapidus SK, Olson JC, Onel K, et al. Health-related quality of life and its relationship to patient disease course in childhood-onset systemic lupus erythematosus. J Rheumatol 2009;36:1536–45.
20. Varni JW, Burwinkle TM, Seid M. The PedsQL 4.0 as a school population health measure: feasibility, reliability, and validity. Qual Life Res 2006;15:203–15.
21. Varni JW, Burwinkle TM, Seid M, Skarr D. The PedsQL 4.0 as a pediatric population health measure: feasibility, reliability, and validity. Ambul Pediatr 2003;3:329–41.
22. Brunner HI, Klein-Gitelman MS, Miller MJ, Trombley M, Baldwin N, Kress A, et al. Health of children with chronic arthritis: relationship of different measures and the quality of parent proxy reporting. Arthritis Rheum 2004;51:763–73.
23. Varni JW, Seid M, Smith Knight T, Burwinkle T, Brown J, Szer IS. The PedsQL in pediatric rheumatology: reliability, validity, and responsiveness of the Pediatric Quality of Life Inventory Generic Core Scales and Rheumatology Module. Arthritis Rheum 2002;46:714–25.
24. Moorthy LN, Peterson MG, Baratelli MJ, Hassett AL, Lehman TJ. Preliminary cross-cultural adaptation of a new pediatric health-related quality of life scale in children with systemic lupus erythematosus: an international effort. Lupus 2010;19:83–8.
25. Moorthy LN, Peterson MG, Baratelli M, Harrison MJ, Onel KB, Chalom EC, et al. Multicenter validation of a new quality of life measure in pediatric lupus. Arthritis Rheum 2007;57:1165–73.
26. Moorthy LN, Peterson MG, Harrison MJ, Onel KB, Lehman TJ. Quality of life in children with systemic lupus erythematosus: a review. Lupus 2007;16:663–9.
27. Moorthy LN, Peterson MG, Hassett AL, Baratelli M, Chalom EC, Hashkes PJ, et al. Relationship between health-related quality of life and SLE activity and damage in children over time. Lupus 2009;18: 622–9.

Summary Table for Measures of Health-Related Quality of Life in Pediatric Systemic Lupus Erythematosus*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
Childhood Health Assessment Questionnaire (C-HAQ)	Measure of disability/ health status	Child/parent self- or interviewer-administered	<10 minutes to complete	<10 minutes to administer; <5 minutes to score	Range 0–3; higher scores indicate higher disability	Excellent internal consistency (Cronbach's α = 0.96)	Moderate correlation with the CHQ PhS, parent's global assessment of pain and overall well-being, and SLAM	Moderate SRM of 0.74	Easy to administer; widely used; excellent responsiveness over time	Limited to assessing physical disability; significant floor effect with less active SLE
Child Health Questionnaire (CHQ)	Generic measure of pediatric HRQOL	Self-administered by child or parent	50-item CHQ Parent Form takes 10–15 minutes to complete	Not reported	Scores for each subscale range from 0–100; higher scores indicate better health status; mean ± SD score is 50 ± 10	Not specifically tested in pediatric SLE	CHQ PhS (physical domain) score correlates well with measures of disease activity and damage; CHQ PsS (psychosocial domain) score only correlates with disease activity	The CHQ PhS changes with disease activity while CHQ PsS does not; SRM for both is <0.4	Comprehensive measure for assessing HRQOL; summary scores are useful; CHQ PhS subscale may be useful in predicting increased disease activity	CHQ PsS does not discriminate between levels of disease activity or damage
Pediatric Quality of Life Inventory Generic Core Module (PedsQL-GC)	Generic measure of pediatric HRQOL	Self-administered by child or parent	<4 minutes to complete	<10 minutes to administer; reportedly easy to score	Summary score ranging from 0–100; higher scores indicate higher HRQOL	Not specifically tested in pediatric SLE	Correlates with BILAG and SDI, but not with SLEDAI or PGA	No significant change in the PedsQL-GC with change in disease activity over time; SRM <0.4	Brief and easily administered	Inconsistent correlation with measures of disease activity; limited ability to assess disease activity over time
Pediatric Quality of Life Inventory Rheumatology Module (PedsQL-RM)	Pediatric rheumatology–specific measure of HRQOL	Self-administered by child or parent	<4 minutes to complete	<10 minutes to administer	Summary score ranging from 0–100; higher scores indicate higher HRQOL	Not specifically tested in pediatric SLE	Correlates with BILAG and SDI, but not with SLEDAI or PGA	Significant change with change in health status; SRM <0.4	Brief and easily administered; correlates with changes in health status over time	May not distinguish well between higher levels of disease activity
Simple Measure of Impact of Lupus Erythematosus in Youngsters (SMILEY)	Pediatric SLE–specific measure of HRQOL	Self-administered by child or parent	<10 minutes to complete; fifth-grade reading level	10 minutes to administer; <10 minutes to score	Total score is transformed to a 1–100 scale; higher scores reflect higher HRQOL	High interrater reliability (ICC 0.7–0.9); high internal consistency (Cronbach's α = 0.9 for both child and parent scores)	Moderate correlation of the total child and parent score with C-HAQ, global quality of life, PedsQL-GC, and domains of the PedsQL-RM; child SMILEY limitation domain had mild correlation with PGA, SLEDAI, and SDI	Changes in the child report SMILEY and particularly the burden of SLE domain correlated best with measures of disease activity and damage over time	Only disease-specific measure for pediatric SLE; excellent evidence of reliability and validity	Total child and parent scores may not adequately predict change in disease activity over time

* PhS = physical health; SLAM = Systemic Lupus Activity Measure; SRM = standardized response mean; SLE = systemic lupus erythematosus; HRQOL = health-related quality of life; PsS = psychosocial health; BILAG = British Isles Lupus Assessment Group index; SDI = Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index; SLEDAI = Systemic Lupus Erythematosus Disease Activity Index; PGA = physician's global assessment; ICC = intraclass correlation coefficient.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S263–S286
DOI 10.1002/acr.20579
© 2011, American College of Rheumatology

Measures of Fatigue

Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire (BRAF MDQ), Bristol Rheumatoid Arthritis Fatigue Numerical Rating Scales (BRAF NRS) for Severity, Effect, and Coping, Chalder Fatigue Questionnaire (CFQ), Checklist Individual Strength (CIS20R and CIS8R), Fatigue Severity Scale (FSS), Functional Assessment Chronic Illness Therapy (Fatigue) (FACIT-F), Multi-Dimensional Assessment of Fatigue (MAF), Multi-Dimensional Fatigue Inventory (MFI), Pediatric Quality Of Life (PedsQL) Multi-Dimensional Fatigue Scale, Profile of Fatigue (ProF), Short Form 36 Vitality Subscale (SF-36 VT), and Visual Analog Scales (VAS)

SARAH HEWLETT, EMMA DURES, AND CELIA ALMEIDA

INTRODUCTION

Fatigue is common to all the rheumatic conditions, in varying degrees, and is a frequent, often severe problem that has major consequences on patients' lives (1–4). In response to these concerns, a body of research subsequently led to international consensus that fatigue must be evaluated in all clinical trials of rheumatoid arthritis and potentially all fibromyalgia syndrome trials (5,6). The 12 fatigue patient-reported outcome measures (PROMs) reviewed in this section have been selected because they are currently or have recently been used in rheumatology populations. Fatigue PROMs in rheumatology were identified from previous reviews (7), then Medline, Cumulative Index to Nursing and Allied Health Literature, and PsycINFO searched for each PROM name plus each major rheumatologic condition. Not all articles could be evaluated and reported in this overview; therefore, those that evidenced strengths and weaknesses were included where possible. However, a full systematic review with meta-analysis would be welcome, as a limitation of this overview is that some articles contributing useful data may have been omitted. The fatigue PROMs are reviewed in alphabetical order. Three additional scales with fatigue components are reviewed elsewhere in this edition: the Bath Ankylosing Spondylitis Disease Activity Index in the Measures of Ankylosing Spondylitis article, the Fibromyalgia Impact Questionnaire in the Measures of Fibromyalgia article, and the Nottingham Health Profile in the Adult Measures of General Health and Health-Related Quality of Life article.

When selecting a fatigue PROM, researchers and clinicians should consider whether their needs are best served by a single-item PROM as a screening tool, by multi-item PROMs that explore broader fatigue issues to create a global score, or by multidimensional PROMs that produce subscale scores for a range of different facets or domains of fatigue (e.g., cognitive and physical fatigue). Multidimensional PROMs with subscales may be useful for informing or evaluating interventions or exploring fatigue causality. Some fatigue PROMs relate to severity only, while others include items of both severity and consequence or impact.

Fatigue PROMs should differentiate between rheumatology populations and healthy controls. Many studies have shown that association between fatigue PROMs and inflammatory markers is not strong, and that fatigue is likely to have multicausal pathways of clinical variables (e.g., pain, disability) and psychosocial variables (e.g., mood, beliefs) combined in varying amounts (1,8–10). Fatigue PROMs should therefore show moderate correlation ($r = 0.3–0.49$) or large correlation ($r = >0.5$) with these variables (11). Very strong associations (e.g., >0.75) might be expected when examining criterion validity with other fatigue scales. Fatigue in rheumatologic conditions can be constant and persistent, but can also appear without warning as an overwhelming event (2–4). Reliability of fatigue PROMs can therefore be problematic to evaluate due to the fluctuating and unpredictable nature of fatigue itself. Some

Dr. Hewlett was in receipt of an unrestricted educational grant from GlaxoSmithKline, which partly-funded a PhD student to develop a rheumatoid arthritis fatigue questionnaire.

Sarah Hewlett, FRCN, PhD, MA, RN, Emma Dures, PhD, MSc, BSc, C Psychol, Celia Almeida, MSc, BSc, BA: University of the West of England, Bristol, UK.

Address correspondence to Sarah Hewlett, FRCN, PhD, MA, RN, Arthritis Research UK Professor of Rheumatology Nursing, University of the West of England, Academic Rheumatology, Bristol Royal Infirmary, Bristol BS2 8HW UK. E-mail: Sarah.Hewlett@uwe.ac.uk.

Submitted for publication April 6, 2011; accepted in revised form July 20, 2011.

fatigue PROMs have therefore been tested for stability over several weeks, and some over a matter of hours, both attempting to capture patients during a stable episode. Test–retest correlations of ≥0.7 are considered acceptable (12). Evaluation data are presented for rheumatology populations, but where these could not be found, data are presented from the original condition in which the PROM was developed.

BRISTOL RHEUMATOID ARTHRITIS (RA) FATIGUE MULTI-DIMENSIONAL QUESTIONNAIRE (BRAF MDQ)

Description

Purpose. The BRAF MDQ was developed to assess the overall experience and impact of RA fatigue, and its different dimensions. It was published in 2010 (13,14).

Content. The BRAF MDQ covers domains of physical fatigue (e.g., average fatigue level over last 7 days), living with fatigue (e.g., has fatigue made it difficult to bathe or shower?), cognitive fatigue (e.g., has fatigue made it difficult to concentrate?), and emotional fatigue (e.g., has being fatigued upset you?).

Number of items. 20 items, providing a total fatigue score, including 4 subscale scores for physical fatigue (4 items), living with fatigue (7 items), cognitive fatigue (5 items), and emotional fatigue (4 items).

Response options. Four options from “Not at all,” “A little,” “Quite a bit,” to “Very much,” except for the first 3 items, which are numerical or categorical as appropriate (e.g., how many days did you experience fatigue in the past 7 days? 0–7).

Recall period for items. The past 7 days.

Endorsements. None found for rheumatologic conditions.

Examples of use. As this is a recently developed patient-reported outcome measure (PROM), there are no additional published studies yet. The BRAF MDQ is currently being used in up to 11 clinical or research studies internationally (information available from the developers).

Practical Application

How to obtain. Available from the developers by e-mail (Sarah.Hewlett@uwe.ac.uk), the web site (available at URL: http://hls.uwe.ac.uk/research/Default.aspx?pageid=312), or by postal mail (Sarah Hewlett, Academic Rheumatology, Bristol Royal Infirmary, Bristol BS2 8HW, UK). The BRAF MDQ is free to use.

Method of administration. Patient self-report, pen and paper.

Scoring. Items scored 0–3, except for items 1 (scored 0–10), 2 (scored 0–7), and 3 (scored 0–2). A total fatigue score is obtained by summing the 20 item scores. Subscale items are summed to produce scores for physical fatigue, living with fatigue, cognitive fatigue, and emotional fatigue. Instructions for missing data are that only 3 questions may be omitted in total, questions 1 and 2 must be completed, and only 1 question may be omitted from each subscale (replaced with patient’s average score for that subscale). Scoring instructions and template can be downloaded from developers’ web site.

Score interpretation. Higher scores reflect greater severity. Total fatigue score is 0–70; subscale scores are physical fatigue 0–22, living with fatigue 0–21, cognitive fatigue 0–15, and emotional fatigue 0–12. In terms of normative data, in the developmental study, 229 people with RA recruited with fatigue visual analog scale (VAS) ≥5 out of 10 had a mean ± SD total fatigue score of 38.4 ± 13.7 of 70, physical fatigue 16.7 ± 4 of 22, living with fatigue 9.6 ± 5.5 of 21, cognitive fatigue 6.1 ± 3.7 of 15, and emotional fatigue 5.8 ± 3.4 of 12 (14). No data for healthy controls could be found.

Respondent burden. Time to complete not reported but probably 4–5 minutes. Items do not appear difficult and have undergone cognitive interviewing.

Administrative burden. Time to score not reported, probably 2–3 minutes using template.

Translations/adaptations. Translated using appropriate linguistic methodology of forward translation, independent back translation by several native speakers, consolidation, then independent back translation to consolidate the final version (information available from the developers). English (UK), French (Belgium), Dutch (Belgium), Spanish (US), German, English (US), Japanese, South Korean, and Taiwan (Chinese) versions can be downloaded from developers’ web site. Over 20 further translations are in progress.

Psychometric Information

Method of development. Items were generated from qualitative research with patients (2), in collaboration with a patient research partner, refined through focus groups, then 45 draft items tested for clarity by cognitive interviewing (13). The 20-item MDQ and its subscales evolved from iterative rounds of Cronbach’s alpha (internal consistency) and factor analysis, with the weakest item removed each time. The resulting 20-item, 4-factor structure was confirmed by a second set of factor analysis on 20 separate, random samples of 50% of the data (bootstrapping) and showed no overlapping items. Subscale labels were discussed and agreed with patient collaborators (14).

Acceptability. The BRAF MDQ appears easily readable; items have undergone cognitive interviewing, with fatigue specifically mentioned in every item. Levels of missing data are not reported. Floor effects (patients unable to report fatigue deterioration) and ceiling effects (patients unable to report improvement) are unlikely to be significant: in 229 patients with RA, <1% scored the maximum possible score for total fatigue, 2.7% for cognitive and for living with fatigue, 4.5% for physical fatigue, and 7% for emotional fatigue; no patients scored the minimum possible fatigue score for total and physical fatigue, <1% for living with fatigue, and 5% and 6% for cognitive and emotional fatigue (in patients recruited with a fatigue VAS >5 out of 10) (15).

Reliability. *Internal consistency.* Cronbach’s alpha for total fatigue was 0.93, physical fatigue 0.71, living with fatigue 0.91, cognitive fatigue 0.92, and emotional fatigue

0.89 (229 patients with RA) (14). Correlations between total fatigue and the 4 subscales range from r = 0.75–0.88 (14).

Test–retest. As fatigue onset is unpredictable and sudden (2), test–retest was conducted 1–2 hours apart (n = 50 patients with RA before and after clinic visits); total fatigue correlated r = 0.95, physical fatigue r = 0.94, living with fatigue r = 0.89, cognitive fatigue r = 0.89, and emotional fatigue r = 0.92 (16).

Validity. *Content validity.* Items and their wording cover a range of fatigue severity and impact and were derived from patient interviews, then refined with focus groups (13).

Construct validity. In 229 patients with RA, total fatigue correlated positively with depression, anxiety, disability, and helplessness (0.50–0.63); subscale physical fatigue (severity) correlated moderately with disability, depression, and helplessness (0.37–0.45), and weakly with anxiety (0.26); living with fatigue correlated positively with depression, anxiety, disability, and helplessness (0.45–0.61); cognitive fatigue correlated moderately with depression, anxiety, and helplessness (0.33–0.49), and weakly with disability (0.21); emotional fatigue correlated positively with depression and anxiety (0.54 and 0.57, respectively) and moderately with helplessness and disability (0.45 and 0.35, respectively); neither total fatigue nor the subscales are strongly associated with pain (0.14–0.38) (14).

Criterion validity. In 229 patients with RA, total fatigue correlated very strongly with the Multi-Dimensional Assessment of Fatigue (RA specific) at 0.82, and the Functional Assessment of Chronic Illness Therapy fatigue subscale at −0.81, and positively with the Short Form 36 vitality subscale (SF-36 VT) at −0.64 (14). A range of moderate to strong correlations were also seen between these measures and the BRAF subscales: physical fatigue (severity) −0.68 to 0.83, living with fatigue −0.54 to −0.74, emotional fatigue −0.50 to 0.66, and cognitive fatigue −0.40 to 0.55 (14). Lower levels of association seen in cognitive fatigue reflect the lack of cognitive fatigue items in other fatigue measures. For total fatigue and all subscales, correlation with SF-36 VT is weaker than with other fatigue scales (see section on SF-36 VT).

Ability to detect change. In patients with RA in flare receiving an intramuscular injection of glucocorticoids (n = 42), effect sizes of 0.33–0.56 for the total BRAF MDQ and subscales were seen at 2 weeks (all $P < 0.04$) (17).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BRAF MDQ is RA specific, and was developed in collaboration with patients, with cognitive interviewing of draft items, and includes the word "fatigue" in every item. Factor analysis shows novel subscales of emotional, cognitive, and living with fatigue, which may help elucidate different causal or perpetuating mechanisms, or highlight individual patient dimensions that require targeted interventions. Internal consistency, test–retest reliability, and construct validity are good, and the BRAF MDQ shows criterion validity with other fatigue scales.

Caveats and cautions. Sensitivity data are still under peer review, and the full article on reliability and sensitivity is awaited. As a recent PROM, it has not yet been widely used; therefore, all evidence is from the developers' article only.

Clinical usability. The available data suggest the BRAF MDQ may be a useful tool in identifying different types of RA fatigue, which might inform individualized self-management interventions. There is no significant administrative or respondent burden.

Research usability. The available data suggest the BRAF MDQ may be a useful research tool in identifying the overall fatigue experience and different types of RA fatigue, and potentially how these might have different causal factors or treatment responses.

BRISTOL RHEUMATOID ARTHRITIS (RA) FATIGUE NUMERICAL RATING SCALES (BRAF NRS) FOR SEVERITY, EFFECT, AND COPING

Description

Purpose. Lack of standardized NRS and visual analog scales (VAS) for fatigue limits the interpretation of data and researchers often create individual items for individual studies (7); therefore, the aim of the BRAF NRS was to develop standardized NRS for measuring a range of RA fatigue domains: severity, effect on life, and coping ability. The BRAF NRS were published in 2010 (13,14).

Content. 3 single-item NRS on fatigue severity (average level of fatigue), effect (effect fatigue has had on your life), and coping (how well you have coped with fatigue).

Number of items. 3, 1 for each concept.

Response options. Patients circle the NRS from 0–10. Anchors are: for severity, "No fatigue" to "Totally exhausted"; for effect, "No effect" to "A great deal of effect"; and for coping, "Not at all well" to "Very well." Initial test–retest data suggested the lack of specificity in the anchors of the coping NRS caused confusion; therefore, anchors were rephrased as "Not coped at all" to "Coped very well" and are being retested (16).

Recall period for items. During the past 7 days.

Endorsements. None found for rheumatologic conditions.

Examples of use. As a recently developed patient-reported outcome measure (PROM), there are no additional published studies yet. The BRAF NRS are currently being used in up to 11 clinical or research studies internationally (information from the developers).

Practical Application

How to obtain. Available from the developers by e-mail (Sarah.Hewlett@uwe.ac.uk), the web site (available at URL: http://hls.uwe.ac.uk/research/Default.aspx?pageid=312), or by postal mail (Sarah Hewlett, Academic Rheumatology, Bristol Royal Infirmary, Bristol BS2 8HW, UK). The BRAF NRS is free to use.

Method of administration. Patient self-report, pen and paper.

Scoring. Each NRS is scored 0–10.

Score interpretation. Scores range from 0–10 with higher scores reflecting greater problems for severity and effect NRS, but lower scores reflecting greater problems for coping NRS. In terms of normative data, 229 people with RA recruited with a screening fatigue VAS ≥5 out of 10 had a mean ± SD BRAF severity NRS of 6.8 ± 1.8, effect NRS of 6.5 ± 2.2, and coping NRS of 5.7 ± 2.3 (14). No data for healthy controls could be found.

Respondent burden. Time to complete not reported but probably 1 minute for the trio. Items do not appear difficult and have undergone cognitive interviewing.

Administrative burden. Time to score not reported but probably 1 minute.

Translations/adaptations. Translated using appropriate linguistic methodology of forward translation, independent back translation by several native speakers, consolidation, then independent back translation to consolidate the final version (information available from the developers). English (UK), French (Belgium), Dutch (Belgium), Spanish (US), German, English (US), Japanese, South Korean, and Taiwan (Chinese) versions can be downloaded from developers' web site. Over 20 further translations are in progress.

Psychometric Information

Method of development. The topics and wording were generated from qualitative research with patients (2), refined by patient research partner, focus groups, and cognitive interviewing (13).

Acceptability. The BRAF NRS appear easily readable, having been developed with patients and undergone cognitive interviewing. Floor effects and ceiling effects are unlikely to be significant: in 229 patients with RA, 4% scored the worst possible score for severity and effect, and 3% for coping; no patients scored the minimum possible score for severity, 0.4% for effect, and 2.6% for coping (patients were recruited with a fatigue VAS >5 out of 10) (15).

Reliability. *Test–retest.* As fatigue onset is unpredictable and sudden (2), test–retest was conducted 1–2 hours apart (n = 50 patients with RA before and after clinic attendance); severity NRS correlated at r = 0.92, effect r = 0.85, and coping r = 0.62 (16). Coping NRS anchors were subsequently reworded to enhance clarity and are currently being retested (16).

Validity. *Content validity.* The single-item NRS cover aspects of fatigue generated from patient interviews, refined with focus groups (13); fatigue coping is not available as a separate domain in other PROMs (7).

Construct validity. In 229 patients with RA, severity NRS correlated moderately with helplessness, depression, disability, anxiety, and pain (0.31–0.45); effect NRS also correlated moderately with these (0.34–0.49); coping NRS correlated moderately with depression (−0.32 to −0.42), weakly with disability and anxiety (−0.21 to −0.29), and not with pain (−0.08); and there was no association between the NRS and raised plasma viscosity (−0.01 to −0.25) (14).

Criterion validity. In 229 patients with RA, the NRS correlated with fatigue measures Multi-Dimensional Assessment of Fatigue, Functional Assessment Chronic Illness Therapy (Fatigue), and Short Form 36 vitality subscale (SF-36 VT): severity 0.65–0.80, effect 0.65–0.75, and coping 0.37–0.38 (14). Lower levels of association seen in coping NRS reflect the lack of coping items in other fatigue measures. For all NRS, correlations with SF-36 VT are weaker than other fatigue scales (see later SF-36 VT section). Correlation between severity and effect (r = 0.71) is strong, while associations between perceived coping and both severity and effect are weak to moderate (r = −0.235 and −0.352, respectively), suggesting coping is a different concept (14).

Ability to detect change. In patients with RA in flare receiving an intramuscular injection of glucocorticoids (n = 42), effect sizes of 0.47 and 0.46 for the BRAF severity and effect short scales were seen, but no significant change in BRAF coping (17).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BRAF NRS are RA specific, and were developed in collaboration with patients, including cognitive interviewing of items. They differentiate fatigue severity from effect and perceived coping ability. They show good construct and criterion validity, and severity and effect show good test–retest reliability. Identically phrased VAS were tested alongside the NRS (13–16). However, the developers recommend the NRS versions as they can be telephone administered, may be conceptually easier to understand and therefore more accurate (13–16,18), and the BRAF NRS showed stronger construct and criterion validity than VAS versions (14–16).

Caveats and cautions. The full article on reliability and sensitivity is awaited. Reverse scoring of the coping NRS may mean that interpretation of coping scores is not immediately obvious, and it has weaker reliability. As a recent PROM, the NRS have not yet been widely used; therefore, all evidence is from developers' articles only.

Clinical usability. The available data suggest the BRAF NRS may be a useful, quick tool to identify 3 different concepts of RA fatigue, which might inform individualized self-management interventions, with no significant administrative or respondent burden.

Research usability. The available data suggest the BRAF MDQ may be a useful research tool to screen for entry criteria, and to identify different facets of fatigue that might be changed differentially by interventions (e.g., fatigue not reduced but perceived coping and impact improved).

CHALDER FATIGUE QUESTIONNAIRE (CFQ)

Description

Purpose. Sometimes referred to as the Chalder Fatigue Scale, or simply the Fatigue Questionnaire or Scale, the

CFQ was developed to assess disabling fatigue severity in hospital and community populations and was originally published in 1993 with further psychometric evaluation in 2010 (19,20).

Content. Covers physical fatigue (e.g., lack energy, feel weak, less muscle strength, need to rest), and mental fatigue (e.g., concentration, memory).

Number of items. 11 items to produce a global score and 2 domains of physical and mental fatigue.

Response options. 4 options, slightly reworded in the latest evaluation: “Less than usual,” “No more than usual,” “More than usual,” and “Much more than usual” (20).

Recall period for items. In the last month.

Endorsements. None found for rheumatologic conditions.

Examples of use. The CFQ has been used in systemic lupus erythematosus (SLE), primary Sjögren’s syndrome (PSS), rheumatoid arthritis (RA), psoriatic arthritis (PsA), fibromyalgia syndrome (FMS), and upper-extremity or carpal tunnel disorder (21–28), as well as chronic fatigue syndrome (CFS) and chronic widespread pain.

Practical Application

How to obtain. From the developer by e-mail: trudie.chalder@kcl.ac.uk. The CFQ is free to use.

Method of administration. Patient self-report, pen and paper.

Scoring. Items can be scored in 2 ways. The first is on a scale of 0–3, giving a global score range of 0–33, a physical fatigue domain range of 0–21 (items 1–7), and a mental fatigue domain range of 0–12 (items 8–11). Second, the CFQ can be scored in a binary fashion (0,0,1,1), then summed to produce a global score of 0–11. No information could be found on handling missing items.

Score interpretation. Higher scores reflect greater fatigue. For Likert scoring, a score of 29 of 33 discriminates clinically relevant fatigue from nonclinically relevant fatigue (20), and for binary scoring, a global score of ≥4 of 11 designates a “case” of fatigue (19). In terms of normative data, mean ± SD Likert scores in a community population (n = 1,615) were 14.2 ± 4.6 of 33 versus 24.4 ± 5.8 in patients with CFS (n = 361; $P < 0.0001$) (20). In FMS, mean ± SD score was 18.2 ± 6.1 (n = 30), and in PsA 6.8 ± 2.8 (n = 9) (21). Mean ± SD binary scores in a community population (n = 1,615) were 3.27 ± 3.21 of 11 versus 9.14 ± 2.73 in 274 patients with CFS (20). Many studies use the draft 14-item CFQ (score range 0–42), giving a mean global fatigue score in 120 patients with SLE of 22 (interquartile range [IQR] 16–28) (22). Using the draft 14-item CFQ with a 5th response option added (score range 0–56), median global fatigue in 51 patients with PSS was 37 (IQR 32–42) versus 28 (IQR 28–32) in 51 controls ($P = 0.000$) (23).

Respondent burden. Time to complete not reported but probably 2–3 minutes. Items appear easy to interpret.

Administrative burden. Time to score not reported, probably 2–3 minutes.

Translations/adaptations. The CFQ comprises 11 items scored 0–33 (19) and underwent minor wording change in 2010 (20). However, while the 8 rheumatology studies identified here quote the original validation article (19), only 2 use this version (21,28). Three articles use the 14-item draft CFQ (22,24,25) giving scores of 0–42; it is not clear from 2 articles which version has been used (26,27), and a Swedish version combined the 14-item draft CFQ with an additional 5th response option (“Much better than usual”), giving a global score of 0–56 (23).

Psychometric Information

Method of development. Fourteen draft items were generated by professionals to represent physical and mental fatigue, and evaluated in new registrants at a general practice (GP; family doctor; n = 274, ages 18–45 years) (19). Factor analysis identified 3 items for removal (19). The resultant 11-item scale includes 2 clear domains on factor analysis (physical fatigue, mental fatigue) with slight overlap between factors for 1 item (concentration) (19,20).

Acceptability. Items appear easy to read. No data on missing item rate or floor/ceiling effects in rheumatology could be found.

Reliability. *Internal consistency.* Cronbach’s alpha was calculated in 274 GP patients for all 14 draft items and by taking out different items one at a time (0.88–0.90) and for the 2 domain scores (physical 0.84, mental 0.82) (19). For the final 11-item version, Cronbach’s alpha was 0.89 in GP patients (n = 274), 0.92 in patients with CFS (n = 361), and 0.88 in a survey of GP attenders (n = 1,615) (19,20). No internal consistency data for rheumatology could be found.

Test–retest. No test–retest data could be found.

Validity. *Content validity.* Items were generated by experts and the final 11-item CFQ covers a range of physical and mental fatigue issues and produces a domain score for each (19).

Construct validity. A CFQ score of 29 of 33 discriminates patients with CFS from the general population in 96% of cases (20). In SLE, using the 14-item draft CFQ, mean fatigue was significantly different between patients (23.5, SEM 0.9; n = 93) and controls (15.0, SEM 0.6; n = 41) (24). However, no difference in total CFQ, physical or mental domains was found between controls and patients with SLE or PSS, and patients with RA only differed from controls in physical fatigue ($P < 0.05$) (28). In SLE, the draft 14-item CFQ scores were moderately associated with each of 4 disease activity measures (r = 0.36–0.4), and with aerobic capacity (r = −0.33) (22). In chronic upper-extremity pain (n = 73), CFQ was moderately associated with pain disability (r = 0.44) and pain intensity (r = 0.32) (27).

Criterion validity. Using a validated psychiatric fatigue interview schedule as a comparator, the cut off for a “case” of fatigue was identified as ≥4 of 11 on the 14-item draft CFQ with 75.5% sensitivity and 74.5% specificity (100 consecutive GP attenders) (19). In SLE (n = 120), the 14-item draft CFQ was strongly associated with the Fatigue Severity Scale and a fatigue visual analog scale (VAS; both r = 0.6) (22). In a group of patients with PSS, RA, and SLE, CFQ total and physical fatigue scores correlated moderately with a fatigue VAS (r = 0.42 and r = 0.46, respectively), but neither the total, physical, nor mental CFQ scores correlated significantly with Short Form vitality subscale (which also did not correlate with the fatigue VAS) (28).

Ability to detect change. In 93 patients with SLE randomized to exercise, relaxation, or control, CFQ improved significantly at exit (22 to 15 versus 24 to 21), and was significantly different between the 8 patients who continued to exercise at 3 months and the 25 who had stopped (11, SEM 5–17 versus 17, SEM 12–26) (26).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CFQ is a fatigue severity scale rather than a measure of impact or consequence, and has physical and mental domains. CFQ has good internal consistency in CFS populations, and good sensitivity to change in rheumatology.

Caveats and cautions. Users need to obtain the correct version (20) from the developer. Many researchers continue to use the draft 14-item version, which makes interpretation across studies difficult. The response options comprise 1 positive, 1 neutral, and 2 negative responses, which might bias Likert scoring (0–3), although using binary scoring (0/1) to define "cases" resolves this issue. There are few rheumatology data on the 2 domains, nor on internal consistency or test–retest. In one study, CFQ did not differentiate between people with rheumatologic conditions and controls. Construct and criterion validity were only moderate in rheumatology.

Clinical usability. A short scale, potentially useful in clinical situations to measure fatigue severity. No significant administrative or respondent burden.

Research usability. Potentially a short, useful patient-reported outcome measure to measure fatigue severity. No significant administrative or respondent burden.

CHECKLIST INDIVIDUAL STRENGTH (CIS20R AND CIS8R)

Description

Purpose. The CIS was developed to measure several aspects of fatigue in chronic fatigue syndrome (CFS) in 1994 (29).

Content. The CIS covers domains of the subjective fatigue experience (e.g., Physically I feel exhausted), concentration (e.g., Thinking requires effort), motivation (e.g., I don't feel like doing anything), and physical activity levels (e.g., I think I do very little in a day).

Number of items. 20 items providing a total CIS20R score, including 4 subscale scores for subjective fatigue experience (8 items), concentration (5 items), motivation (4 items), and physical activity levels (3 items). Although the entire CIS20R assesses fatigue, the 8-item subjective fatigue subscale is commonly the only subscale reported and is often referred to as CIS8R, CIS-Fatigue, or Fatigue Severity.

Response options. 7 boxes ranging from "Yes that is true," to "No that is not true."

Recall period for items. The past 2 weeks.

Endorsements. None found for rheumatologic conditions.

Examples of use. Mainly used in CFS, multiple sclerosis, neurologic disorders, and healthy working adults, but has been used in rheumatoid arthritis (RA) and fibromyalgia syndrome (FMS) (30–39).

Practical Application

How to obtain. From the developer by e-mail: j.vercoulen@mps.umcn.nl. The CIS is free to use.

Method of administration. Patient-self-report, pen and paper.

Scoring. Items scored 1–7, with 11 positively phrased items reverse scored. A total CIS20R score is obtained by summing the 20-item scores. Subscale items are summed to produce scores for subjective fatigue (CIS8R), concentration, motivation, and physical activity. No information is given on handling missing items.

Score interpretation. Higher scores reflect greater severity. Overall CIS20R score is 20–140; subscale scores are subjective fatigue (CIS8R) 8–56, concentration 5–35, motivation 4–28, and physical activity 3–21. In terms of normative data, in healthy controls ($n = 60$), mean ± SD subjective fatigue was 2.4 ± 1.4, concentration 2.2 ± 1.2, motivation 2.0 ± 1.0, and physical activity 2.0 ± 1.3 (29). Cut offs on the subjective fatigue CIS8R scale for patients with RA are based on the mean score for healthy adults plus 1 or 2 SDs, i.e., 27–35 for heightened fatigue, and ≥35 for severe fatigue (31), with ≥35 reported as similar to fatigue levels in CFS (38). In patients with RA ($n = 228$), mean ± SD subjective fatigue CIS8R was 31.5 ± 12.8, with 20% reporting heightened fatigue and 42% reporting severe fatigue (31).

Respondent burden. Time to complete not reported but probably 4–5 minutes.

Administration burden. Time to score not reported, probably 4–5 minutes to reverse score some items, then identify and sum subscale items.

Translations/adaptations. The CIS originates from The Netherlands. Dutch, English, Swedish, and Korean versions are available from the developer.

Psychometric Information

The subjective fatigue subscale (CIS8R) is the most commonly and often the only reported data in studies using the CIS.

Method of development. No information could be found on how items were generated; 20 of the original 24 draft items were retained as they performed best in factor analysis (29). Subscales were generated through principal components analysis and Cronbach's alpha (internal consistency) (29). Evaluation in RA is reported in an abstract (37).

Acceptability. In a rheumatology population, 3 items might be interpreted in relation to RA or disability ("I feel fit," "Physically I feel I am in bad form," "Physically I feel I am in an excellent condition") and thus may not be sensitive to RA fatigue. Levels of missing data and floor/ceiling effects are not reported.

Reliability. *Internal consistency.* In CFS, total CIS20R score Cronbach's alpha was 0.90 and Gutman split-half reliability coefficient 0.92; Cronbach's alpha for subscales ranged from 0.83–0.92 (29). In patients with RA (n = 227), Cronbach's alpha for subjective fatigue CIS8R was 0.92 (37), and 0.89 in patients with FMS (n = 78) (36). In patients with RA, factor analysis is reported as confirming the 4 subscales (no data are provided) (n = 227) (37).

Test–retest. In 227 patients with RA, intraclass correlation coefficient for subjective fatigue CIS8R over 1 month was 0.81 (37).

Validity. *Content validity.* No information is provided on how items were generated (29) but the CIS20R covers a range of fatigue issues likely to be common in rheumatology populations (2–4).

Construct validity. In patients with RA (n = 228), subjective fatigue CIS8R correlated strongly with pain (0.55), moderately with disability, sleep disturbance, helplessness, anxiety, and depression (0.32–0.40), weakly with rheumatoid factor, Disease Activity Score in 28 joints, and tender or swollen joints (0.18–0.3), and not with disease duration or inflammatory indices (31). The total CIS20R score discriminates between healthy workers and workers with health reasons for being fatigued (39).

Criterion validity. In patients with RA (n = 227), subjective fatigue CIS8R correlated very strongly with Short Form 36 vitality subscale and with a fatigue numerical rating scale (both 0.81) (37). In patients with FMS (n = 224), subjective fatigue CIS8R correlated with a fatigue visual analog scale (VAS) at 0.61 (35).

Ability to detect change. In patients with FMS (n = 78) receiving cognitive–behavioral therapy (CBT), subjective fatigue CIS8R improved by a mean ± SD −10.6 ± 10.7 (36). In patients with RA started on anti–tumor necrosis factor therapy (n = 126), total CIS20R score improved from a mean 85 (65–97) to 69 (48–90) over 6 months (30), while in a subset of 59 working-age patients, CIS20R score improvement was 11.8% at 6 months (33). In early, distressed patients with RA (n = 30), CBT gave an effect size of 0.55 posttreatment for subjective fatigue CIS8R (0.48 at 6 months) (34). No minimum clinically important difference is reported, but in FMS (n = 78), change in subjective fatigue CIS8R correlated with a transition question on perceived change (0.53), and with a VAS for usefulness of and satisfaction with the level of change (0.42 and 0.33, respectively) (36).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CIS was developed in CFS but has been evaluated in many long-term conditions, suggesting it is a useful generic scale. The CIS has good internal consistency and reliability, construct and criterion validity, and sensitivity to change. Subscales differentiate between cognitive and physical fatigue.

Caveats and cautions. The full article evaluating use of CIS in RA is awaited. Most evidence is reported only for the subjective fatigue subscale. Three items may be confounded by disability or disease activity in rheumatology populations.

Clinical usability. The available data suggest the CIS may be a useful tool in identifying cognitive and physical fatigue, which might inform individualized self-management interventions. No significant respondent or administrative burden.

Research usability. The available data suggest the CIS may be a useful research tool in identifying both the overall fatigue experience and different types of fatigue, within the caveats above.

FATIGUE SEVERITY SCALE (FSS)

Description

Purpose. The FSS was developed to assess disabling fatigue in multiple sclerosis (MS) and systemic lupus erythematosus (SLE), and was published in 1989 (40).

Content. The FSS covers physical, social, or cognitive effects of fatigue (e.g., function, work, motivation).

Number of items. 9 items to produce a global score.

Response options. 7 options from "Strongly disagree" to "Strongly agree" (1–7).

Recall period for items. The past week.

Endorsements. After systematic review of 15 fatigue scales used in SLE, the Ad Hoc Committee recommended the FSS for use in SLE (41).

Examples of use. Has been used extensively in SLE studies, and also in rheumatoid arthritis (RA), osteoarthritis (OA), and ankylosing spondylitis (AS), with a modified version in psoriatic arthritis (PsA) (41–54), as well as many long-term conditions (e.g., MS, cancer, neurologic disorders).

Practical Application

How to obtain. From the developer by e-mail: lkrupp@notes.cc.sunysb.edu. The FSS is free to use.

Method of administration. Patient self-report, pen and paper.

Scoring: Items are scored 1–7, summed, then averaged to produce a global score.

Score interpretation. Scores range from 1–7 with higher scores reflecting greater fatigue. In terms of normative data, mean ± SD score in healthy adults (n = 20) was 2.3 ± 0.7, compared to 4.7 ± 1.5 in patients with SLE, and 4.2 ± 1.2) in patients with RA (n = 29 and 122, respectively) (40,44). In patients with PsA (n = 135) using a modified FSS scaled from 0–10 (mFSS; see adaptations below), mean score was 5.7 (95% confidence interval [95% CI] 5.1–6.3) (45). In another PsA study (n = 75) using the mFSS, patients reporting fatigue in a clinical assessment had a mean ± SD mFSS score of 6.9 ± 2.4 compared to 3.8 ± 2.8 in those reporting no fatigue (42). In OA (n = 137), mean ± SD FSS was 3.63 ± 1.55 (54).

Respondent burden. Time to complete not reported but probably 2–3 minutes. Items appear easy to interpret.

Administrative burden. Time to score not reported, probably 2–3 minutes.

Translations/adaptations. Translated into multiple languages, including Spanish, French, Chinese, and Portuguese (41) with a Swedish translation describing appropriate linguistic methodology, then evaluation of reliability, and construct and criterion validity in SLE (46). Adaptations include a multidimensional, 29-item Fatigue Assessment Instrument in German (55); a US adaptation for telephone administration in RA, which reduced the response options from 1–7 to 1–5 and states FSS has 10 rather than 9 items (47); and an mFSS used in PsA that increased the response options from 1–7 to 0–10 ("Not at all" to "Entirely"), although no rationale for this was presented (45).

Psychometric Information

Method of development. Factor analysis was performed on 28 draft items, and identified 9 items common to both SLE and MS (n = 29 and 25, respectively); it is not stated where the 28 items originated from or whether patients were involved in their development (40).

Acceptability. Items appear easy to read, the Swedish version underwent cognitive debriefing (46) and fatigue is mentioned in every question. In one study, none of the 22 patients with SLE omitted any questions; the study reported no ceiling effects, but a possible floor effect for 1 item, where the median score was also the maximum possible score (46).

Reliability. *Internal consistency.* Cronbach's alpha was 0.89–0.94 in SLE (n = 22–29) (40,46). In the mFSS (0–10 response option), Cronbach's alpha was 0.95 in both PsA (n = 91) and SLE (n = 113) (45).

Test–retest. No significant difference was seen in FSS in stable patients with SLE over 1 week (46).

Validity. *Content validity.* FSS covers a range of fatigue issues. It is not stated how items were generated (40) but the FSS later underwent cognitive testing in Swedish patients with SLE (46).

Construct validity. FSS correctly discriminated 90% of 29 patients with SLE from healthy controls (40). A systematic review reports evaluation of construct validity of the FSS in a number of SLE studies, demonstrating a range of correlations with disease activity (0.16–0.53), depression (0.22–0.59), and pain (0.35–0.54) (41). In patients with SLE (n = 22), FSS correlated strongly with pain, general health, and physical and social roles (−0.59 to −0.60), and moderately with function, emotional role, and mental health (−0.41 to −0.44) (46); no association was found with inflammatory indices (n = 57 SLE) (48). In an RA working population (n = 122), FSS correlated with anxiety and depression (0.55 and 0.53, respectively), and disability, pain, and stress (0.33–0.48) (43). In PsA (n = 135), the mFSS (10-point response) correlated with number of active joints (0.37), but not swollen or damaged joints (49).

Criterion validity. mFSS correlated strongly with a fatigue visual analog scale at 0.81 (SLE, n = 29) (40) and with Functional Assessment Chronic Illness Therapy (Fatigue) at −0.79 in patients with PsA (n = 135) (49). Correlation with Short Form 36 vitality subscale (SF-36 VT) was −0.56 to −0.63 in SLE, OA, and RA (n = 32, 137, and 52, respectively) (46,54).

Ability to detect change. In patients with AS randomized to etanercept or placebo (n = 40), FSS showed an effect size of 0.43 for treatment at 4 months (SF-36 VT effect size 0.69); FSS was not responsive at 1 month, unlike SF-36 VT (effect size 0.15 versus 0.54) (50). In SLE (n = 58), effect sizes of 0.55 and 0.44 were shown from telephone interventions for fatigue (modified 10-item, 5-response option FSS) (47). Based on linear regression analysis on comparative fatigue ratings from patients after paired interviews, the effect size (mean change/SD at baseline) required for an average patient to move to a different fatigue category (i.e., much, somewhat or a little, less or more fatigued) is calculated as 0.74 in RA (52) and 0.41 (95% CI 0.2–0.57) in SLE, where the authors also present this as an FSS minimum clinically important difference score of 0.6 (95% CI 0.3–0.9) (53). Based on a systematic review of earlier SLE studies, a recommendation for important improvement in FSS for patients with SLE was 15% (41).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The FSS has good internal consistency, reliability, and construct and criterion validity, and is sensitive to change. It has been evaluated in several rheumatologic conditions, particularly SLE, where it is the recommended fatigue scale (41).

Caveats and cautions. Differences in sensitivity compared to SF-36 VT were shown in one study, but without a third fatigue comparator patient-reported outcome measure (PROM) for that study, it is not possible to conclude whether FSS or SF-36 VT is more accurate. Comparison between rheumatologic groups may be difficult if some groups use the differently scaled mFSS rather than the FSS.

Clinical usability. A short scale, potentially useful in clinical situations.

Research usability. Potentially a short, useful PROM for research, although researchers should be aware items suggest FSS may measure fatigue impact rather than severity.

FUNCTIONAL ASSESSMENT CHRONIC ILLNESS THERAPY (FATIGUE) (FACIT-F)

Description

Purpose. The FACIT-F was developed in 1997 to measure fatigue in oncology patients with anemia and is a stand-alone (or add-on) questionnaire in the Functional Assessment in Cancer Therapy measurement system (56). This has since been widened to include assessment of chronic illnesses (FACIT measurement system). The current version of FACIT-F is number 4.

Content. The FACIT-F covers physical fatigue (e.g., I feel tired), functional fatigue (e.g., trouble finishing things), emotional fatigue (e.g., frustration), and social consequences of fatigue (e.g., limits social activity).

Number of items. 13 to produce a global score.

Response options. 5 responses from "Not at all" to "Very much."

Recall period for items. Past 7 days.

Endorsements. None found for rheumatologic conditions.

Examples of use. Has been evaluated in rheumatoid arthritis (RA) and psoriatic arthritis (PsA), and used in primary Sjögren's syndrome (PSS), osteoarthritis (OA), and systemic lupus erythematosus (SLE) (49,52,53,57–63), as well as many long-term conditions (e.g., multiple sclerosis, cancer, neurologic disorders).

Practical Application

How to obtain. From the FACIT web site after free registration at URL: http://www.facit.org/. English versions are free to use, a fee is payable for non-English versions used in commercial studies.

Method of administration. Patient self-report, interviewer or telephone administered.

Scoring. Items scored 0–4, with 2 positively phrased items reverse scored. Items are summed, multiplied by 13, then divided by the number of items actually answered, therefore allowing for missing items. However, more than 50% of items must be answered (i.e., at least 7 items). Scoring instructions can be downloaded from the developers' web site, including computerized versions.

Score interpretation. Scores range from 0–52, with higher scores reflecting less fatigue. In terms of normative data, mean ± SD score for 1,010 healthy adults was 43.6 ± 9.4 (64); this compares to 29.17 ± 11.06 in patients with RA, 35.8 ± 12.4 in patients with PsA, 25.7 ± 12.0 in patients with SLE, and 30.1 in patients with PSS (n = 631, 135, 80, and 277, respectively) (49,53,57,63).

Respondent burden. 3–4 minutes. Items appear easy to interpret.

Administrative burden. Time to score not reported but probably 3–4 minutes.

Translations/adaptations. Available in over 50 languages.

Psychometric Information

Method of development. Items were generated in semistructured interviews with 14 anemic oncology patients and 8 clinicians, followed by item reduction by 5 medical experts, then evaluation in 49 oncology patients (56). In patients with RA (n = 271), analysis using item response theory suggested the FACIT-F covers a wider range of fatigue (with the exception of those with a very low level) than either the Short Form 36 vitality subscale (SF-36 VT) or Multi-Dimensional Assessment of Fatigue (MAF) (57).

Acceptability. The items are brief and easy to understand. However, in arthritis populations, some items have the potential for misinterpretation. Two items could potentially be interpreted as relating to disability rather than fatigue as fatigue is not stipulated in the wording ("Ability" and "Needing help to do usual activities"), 1 item measures energy, which may be a positive health state that is not necessarily the opposite end of a fatigue continuum (i.e., people who are not feeling energized may not necessarily feel fatigued), and 1 item that is applicable to patients with cancer may hold less relevance for patients with RA ("Feeling too tired to eat" is not reported in qualitative RA fatigue studies) (2,3,65). Floor/ceiling effect data could not be found in rheumatology.

Reliability. *Internal consistency.* Cronbach's alpha was 0.86–0.87 at 3 time points in RA (n = 631) and 0.96 in PsA (n = 135) (49,57).

Test–retest. Intraclass correlation coefficient over 1 week was 0.95 in patients with PsA (n = 73) (49).

Validity. *Content validity.* Items were generated by patients with cancer (56) but cover a range of fatigue issues likely to be common to arthritis (2–4).

Construct validity. In PsA (n = 135), FACIT-F correlated with inflamed joint count (r = −0.43, 95% confidence interval [95% CI] −0.56 to −0.28) but not with damaged joint count (r = 0.06, 95% CI −0.23 to 0.11), age, or disease duration (49). In RA (n = 505), FACIT-F correlated with disability (Health Assessment Questionnaire) and inflammation (Disease Activity Score in 28 joints) at r = −0.42 to −0.44 (60). FACIT-F scores were not statistically significantly different between patients with OA (n = 43) and PSS (n = 71), but sleepiness was more strongly associated with FACIT-F in PSS than in patients with OA (0.53 versus 0.27) (58).

Criterion validity. In RA, FACIT-F correlated strongly with MAF at 0, 12, and 24 weeks of antirheumatic treatment (−0.84 to −0.88), and with SF-36 VT (0.73–0.84) (n = 567–631) (57). In PsA (n = 135), correlation with modified Fatigue Severity Scale was −0.79 (95% CI −0.85 to −0.72) while those patients responding positively to an anchor question on overwhelming fatigue had lower FACIT-F scores (i.e., more severe fatigue) than those responding negatively (mean ± SD 24.8 ± 13.9 versus 38.5 ± 10.4; n = 135) (49).

Ability to detect change. After 24 weeks of antirheumatic treatments in patients with RA (n = 631), FACIT-F showed a mean change of 2.1 in patients who did not achieve American College of Rheumatology 20% criteria for improvement in disease activity (ACR20; effect size 0.19), compared to 12.4 in those who achieved ACR70 (effect size 1.13) (57). Sensitivity has also been shown in other anti–tumor necrosis factor trials in RA (61,62), and in PsA where changes in FACIT-F were similar to changes in SF-36 VT (n = 313) (59). A minimum clinically important difference (MCID) of 3–4 points is generally used, which was calculated using 0.2 and 0.5 effect size cut offs for 5 groups ("Major worsening" to "Major improvement") in 631 patients with RA receiving antirheumatic treatments, then confirmed in a second study (n = 271) (57). On a normalized scale of 0–100 (rather than 0–52), others have proposed an MCID for RA of 15.9 points (52). In SLE (n = 80), based on linear regression analysis on comparative fatigue ratings from patients after paired interviews, the effect size required (mean change/SD at baseline) for an average patient to move to a different fatigue category (i.e., much, somewhat or a little, less or more fatigued) is calculated as 0.5 (95% CI 0.31–0.65), which the authors also present as a FACIT-F MCID score of −5.9 (95% CI −8.1 to −3.6) (53).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. FACIT-F is used across many rheumatologic conditions, particularly in pharmacologic trials. It covers a range of fatigue concepts in easy to understand language. FACIT-F has good internal consistency and reliability, construct and criterion validity, and sensitivity to change.

Caveats and cautions. FACIT-F might potentially be limited for use in rheumatology by the phrasing of 4 of the 13 items.

Clinical usability. Would be easy to use in clinical practice, giving a global fatigue score.

Research usability. Would be easy to use in research where a global fatigue score is required.

MULTI-DIMENSIONAL ASSESSMENT OF FATIGUE (MAF)

Description

Purpose. The MAF was developed in 1991 to measure multiple dimensions of fatigue in adults with rheumatoid arthritis (RA) (66). It was a revision of the Piper Fatigue Scale, which had been developed and tested with oncology patients (67).

Content. The MAF covers 4 dimensions of fatigue: severity, distress, interference in activities of daily living (doing chores, cooking, bathing, dressing, working, visiting, sexual activity, leisure, shopping, walking and exercising), and frequency and change during the previous week.

Number of items. 15 items provide a global score (Global Fatigue Index [GFI]). The 16th question ("To what degree has your fatigue changed during the past week?") does not contribute to the GFI.

Response options. The number of response options depends on the nature of each item. The original version used visual analog scales (VAS) for items 1 and 4–14, but based on feedback from respondents, these were changed to numerical rating scales (NRS) ranging from 1–10 in 1995 (68). Items 1 (degree) and 4–14 (interference) have anchors of "Not at all" to "A great deal," item 2 (severity) has anchors of "Mild" to "Severe," and item 3 (distress) has anchors "No distress" to "A great deal of distress." Items 4–14 (interference with activities) provide an opportunity for respondents to indicate if they do not carry out the activity because of reasons other than fatigue, and the item is then not completed. Items 15 and 16 have 4 ordinal response options scored 1–4, with item 15 (frequency) ranging from "Hardly any days" to "Every day," and item 16 (change) ranging from "Decreased" to "Increased."

Recall period for items. The past week.

Endorsements. None found for rheumatologic conditions.

Examples of use. Although developed for use in RA, the MAF has also been used in other rheumatologic conditions, including osteoarthritis (OA), ankylosing spondylitis (AS), systemic lupus erythematosus (SLE), and fibromyalgia syndrome (FMS) (52,53,55,57,68–79), as well as other long-term conditions such as human immunodeficiency virus, multiple sclerosis, and cancer.

Practical Application

How to obtain. The MAF is copyrighted by the developer, Basia Belza, and may be downloaded after free registration from the web site available at URL: www.son.washington.edu/research/maf/, or obtained by postal mail at the following address: Basia Belza, PhD, RN, Department of Biobehavorial Nursing and Health Systems, Box 357266, University of Washington, Seattle, WA 98195-7266. There is no charge for individual use of the MAF, although a nominal fee may be charged for commercial use.

Method of administration. Patient self-report, pen and paper.

Scoring. The MAF was developed to provide an aggregated score, the GFI. If the respondent indicates "No fatigue at all" for item 1, all remaining items should be scored as 0. Items 1–3 are summed, items 4–14 are averaged but should not be scored where the respondent indicates that they do not do an activity "For reasons other than fatigue," and item 15 is transformed into a 0–10 scale by multiplying the score by 2.5. The GFI is then calculated by adding these 3 components (sum of items 1–3, average of items 4–14, and transformed item 15). Item 16 (change) does not contribute to the GFI, and is scored 1–4. No information is given on handling missing data, but the developer has suggested that nonresponse to ≥3 of the 16 items would mean the GFI could not be calculated (Tack BB: unpublished observations).

Score interpretation. The GFI ranges from 1 (no fatigue) to 50 (severe fatigue). A higher score represents greater fatigue severity, distress, or interference with activities of daily living. Item 16 (change) is scored from 1 (fatigue decreased) to 4 (fatigue increased). In terms of normative data, in healthy controls (n = 46), mean ± SD GFI was 17.0 ± 11.3 (68). In rheumatologic conditions, mean ± SD GFI was 29.2 ± 9.9 in RA, 32 ± 20 in AS, 36.4 ± 8.1 in FMS, 31.1 ± 11.4 in SLE, and 27.7 ± 10.8 in OA (n = 51–1,636) (53,68,71,74,76).

Respondent burden. Time to complete is not reported but is likely to be ~5–8 minutes.

Administrative burden. Time to score is not reported but is likely to be ~4–5 minutes to transform scores, sum and average dimensions, and create the GFI.

Translations/adaptations. The MAF was originally developed in US English. The MAPI Research Institute has versions in Spanish, Dutch, French, Mandarin, Croatian, Danish, Finnish, Czech, German, Turkish, Swedish, Afrikaans, Russian, Portuguese, Polish, Italian, Hungarian, Hebrew, and Norwegian. Translation was undertaken using both forward and backward translations. Translations can be obtained through the MAPI web site at URL: http://www.mapi-institute.com/home.

Psychometric Information

Method of development. Items from a 41-item cancer fatigue scale (67) that were considered to describe activities often affected in RA were selected for the MAF (66). No patients were involved in selecting items. Following patient feedback from early studies, the VAS format was changed to NRS in 1995 (68).

Acceptability. Items appear easy to understand, but some may contain overlapping concepts; walking (item 13) and exercise (item 14) might both be considered leisure activities (item 11). Response options for frequency over the past week (item 15) might be subject to different interpretations; for example, when wishing to report 2 days of fatigue, some patients might consider that "Occasionally" and others might consider it "Hardly any days." At the start of items 4–14, respondents are clearly instructed to consider to what degree fatigue has interfered with activities, but fatigue is subsequently not mentioned in the question stems; thus, respondents might inadvertently consider interference due to disability rather than fatigue when scoring these 11 items. Missing data have been reported as a problem; based on the MAF not being able to be scored if they have ≥3 missing items, 2 studies report 21.5% of questionnaires (49 of 229) and 13.9% of questionnaires (1,077 of 7,760) to be unusable (14,70). In RA (n = 271), item response theory suggests the MAF covers the middle range of fatigue severity, broader than Short Form 36 vitality subscale (SF-36 VT) but slightly less than the Functional Assessment Chronic Illness Therapy (Fatigue) (FACIT-F) (57). Floor/ceiling effect data could not be found.

Reliability. *Internal consistency.* Cronbach's alpha for internal consistency was 0.93 in the original VAS version (n = 133 patients with RA), 0.92 for the final NRS version (n = 122 patients with RA), and 0.92 in knee OA (n = 44) (69,77,79).

Test–retest. No significant change in MAF over 3 time points (6–8 week intervals) is reported for patients with RA (n = 51) (68); in cancer (n = 37), test–retest was r = 0.87 over 48 hours (79).

Validity. *Content validity.* The MAF covers a range of fatigue issues (severity, distress, interference with activities, frequency, and change) to create a single, composite score (GFI). The original factor analysis in RA (n = 35) showed that the 15 items comprising the GFI load on a single factor (all >0.55) (66). A later analysis in RA (n = 7,760) indicated 3 factors: interference with leisure-type activities; interference with bathing/dressing; and fatigue frequency, degree, severity, and distress, with 4 further items loading across all 3 factors equally (70).

Construct validity. In RA (n = 51), MAF correlated with depression, pain, disability, and sleep (r = 0.47–0.58) and very weakly with inflammatory markers (0.12) (68). MAF discriminated between people with RA (n = 48) with and without prior history of depression (34.3; SD 10.0 versus 28.8; SD 9.5) (77). In knee OA (n = 44), MAF correlated with female sex, pain, depression, anxiety, and cardiorespiratory stamina (r = 0.52–0.62), but not with muscle (quadriceps) fatigue (r = 0.01) (78). In AS (n = 68), MAF correlated moderately with pain and hemoglobin (0.39 and −0.38, respectively), weakly with SF-36 mental health (−0.27), and weakly but not significantly with SF-36 emotional role (−0.22) (72).

Criterion validity. In RA, MAF correlated strongly with the Profile of Mood States fatigue and vigor subscales at 0.84 and −0.62, respectively (n = 51) (68), with a fatigue VAS at 0.8 (n = 7,760) (70), and with an NRS of "bothersome fatigue" at 0.69 (n = 48) (77). Correlation with SF-36 VT was variable, ranging from −0.79 in RA (n = 7,760) to −0.54 in OA (n = 137) and −0.37 in AS (n = 68) (55,70,72).

Ability to detect change. In RA (n = 631), after 24 weeks of antirheumatic treatments, MAF showed a mean change of −2.1 in patients who did not achieve American College of Rheumatology criteria for 20% improvement in disease activity (ACR20; effect size −0.18), compared to mean change of −14.9 in those who achieved ACR70 (effect size −1.25) similar to findings for FACIT-F and SF-36 VT (57). In FMS (n = 267), after 8 weeks of esreboxetine, MAF improved by −6.39 (SE 0.75) compared to −2.82 (SE 1.74) on placebo (75). Based on linear regression analysis on comparative fatigue ratings from patients after paired interviews, the effect size required (mean change/SD at baseline) for an average patient to move to a different fatigue category (i.e., much, somewhat or a little, less or more fatigued) is calculated as 0.75 in RA (n = 61) (52) and 0.45 (95% confidence interval [95% CI] 0.25–0.61) in SLE (n = 80), where the authors also present this as MAF minimum clinically important difference score of 5.0 (95% CI 2.8–7.2) (53).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The MAF is RA specific and covers numerous aspects of fatigue in order to produce a global score. It has good internal consistency, construct and criterion validity, reliability, and it is sensitive to change.

Caveats and cautions. High levels of missing data are reported, making a substantial proportion of questionnaires unusable. The lack of reference to fatigue on the 11 items asking about interference with activities may reduce clarity for patients who may respond with regards to disability interference.

Clinical usability. The MAF might be useful in clinical practice in providing a global score, while the multiple questions might help identify target areas for therapeutic intervention.

Research usability. The MAF produces a global score based on a range of fatigue impacts. It has reasonable participant and administrative burden, although problems in scoring may arise where there are a large amount of missing items.

MULTI-DIMENSIONAL FATIGUE INVENTORY (MFI)

Description

Purpose. The MFI was originally developed to measure cancer fatigue using a multidimensional, short questionnaire, specifically without any somatic items (80,81). Published in 1995, it was evaluated initially in cancer and chronic fatigue syndrome (CFS) patients and in healthy volunteers who might be physically tired (army recruits) or cognitively tired (junior doctors) (80).

Content. The MFI covers domains of general fatigue (e.g., I feel tired), physical fatigue (e.g., physically I feel only able to do a little), activity (e.g., I feel very active), motivation (e.g., I dread having to do things), and mental fatigue (e.g., my thoughts easily wander).

Number of items. 20 items, yielding 5 subscales of 4 items each (general fatigue, physical fatigue, reduced activity, reduced motivation, and mental fatigue). Creating a total score is discouraged by the developers.

Response options. 5 check boxes ranging from "Yes that is true," to "No that is not true."

Recall period for items. This is stated as "Lately."

Endorsements. None found for rheumatologic conditions.

Examples of use. In addition to cancer and several long-term conditions (e.g., Parkinson's disease, liver disease), MFI has been used in a number of studies in rheumatoid arthritis (RA), fibromyalgia syndrome (FMS), ankylosing spondylitis (AS), primary Sjögren's syndrome (PSS), and systemic lupus erythematosus (SLE) (41,52,53,82–88).

Practical Application

How to obtain. From the developers by e-mail: e.m.smets@amc.uva.nl. Also available by postal mail at the following address: E. M. A. Smets, PhD, Medical Psychology J3-220, Academic Medical Center, University of Amsterdam, PO Box 22660, 1100 DD Amsterdam, The Netherlands. The MFI is free for academic use, charges apply for commercial use.

Method of administration. Patient self-report, pen and paper.

Scoring. Items scored 1–5, with 10 positively phrased items reverse scored. Subscale items summed to produce scores for general fatigue, physical fatigue, reduced activity, reduced motivation, and mental fatigue.

Score interpretation. Scores range from 4–20 with higher scores reflecting greater severity. In terms of normative data, in healthy women (n = 32), general fatigue mean ± SD score was 8.16 ± 3.8 compared to 15.57 ± 4.3 in PSS (n = 49), and 12.93 ± 4.5 in RA (n = 44), physical fatigue 6.47 ± 3.2 versus 14.06 ± 4.4 and 12.45 ± 5.0, reduced activity 6.72 ± 3.0 versus 11.32 ± 4.6 and 11.48 ± 4.7, reduced motivation 6.66 ± 2.4 versus 9.96 ± 4.0 and 9.27 ± 4.1, and mental fatigue 6.53 ± 3.0 versus 10.31 ± 5.4 and 8.34 ± 4.0 (82). In 53 women with FMS, scores were more severe than RA or PSS with subscale mean ± SD of 17.9 ± 1.9, 16.2 ± 2.7, 15.1 ± 3.9, 12.9 ± 3.0, and 14.4 ± 3.5, respectively (84).

Respondent burden. Time to complete not reported but is likely to be 4–5 minutes. Items appear easy to understand.

Administrative burden. Time to score not reported but is likely to be 4–5 minutes to reverse score some items, then identify and sum subscale items.

Translations/adaptations. Authorized translations in most European languages can be obtained from the developers.

Psychometric Information

Method of development. 24 draft items were generated based on existing literature (80) and pilot in-depth interviews with patients with cancer (81), from which the developers postulated the 5 fatigue domains, for each of which they tried to create brief, positively and negatively phrased items that exclude somatic issues (80). Factor analysis on the 24 items supported the 5 subscales with adjusted goodness of fit index (AGFI) ranging from 0.95–0.98 (111 patients with cancer and 357 patients with CFS) (80); the 4 items with the weakest correlations were later removed, leaving a 20-item scale with 4 items per subscale, which had AGFI properties >0.9 (n = 97, 116 patients with cancer) (81). The original MFI-20 had 7 response options (80) but this was revised to the current version with 5 response options following evaluation (81).

Acceptability. The items are brief and easy to understand. Missing item levels are low with 98.2–99.4% completion rates in FMS (n = 166) (87). Items do not contain the word fatigue and thus could potentially be interpreted by rheumatology patients as relating to disability (e.g., I think I do very little in a day) or disease activity (e.g., physically I feel I am in a bad condition). In patients with cancer (n = 116), 10.4–33.6% scored the best possible score for the different subscales (mental fatigue 33.6%), suggesting a potentially substantial ceiling effect; 4.5–15.7% scored the worst possible score (reduced activity 15.7%), suggesting a lesser, but still potentially important floor effect (81). No data could be found for rheumatology populations.

Reliability. *Internal consistency.* Cronbach's alpha for most subscales ranged from 0.85–0.89 in 82 patients with RA or PSS, with reduced motivation at 0.68 (86).

Test–retest. In AS and PSS (n = 40 and 28, respectively), repeat administrations at between 2 and 42 days gave intraclass correlation coefficients (ICCs) of 0.57–0.85 across the subscales (83,84). The ICC in patients with chronic widespread pain or FMS (n = 36) ranged from 0.75–0.92 (87).

Validity. *Content validity.* The MFI covers 5 domains of fatigue, which resonate with qualitative studies in rheumatology (2–4). In PSS, 29 patients scored the coverage of fatigue by the MFI as a mean ± SD 2.96 ± 0.6 on a scale of 1–4 ("Poorly" to "Very well") (84).

Construct validity. All subscales differentiated between fatigued and nonfatigued patients with AS (n = 415 and 361, respectively) based on a cut off of 5 out of 10 on a fatigue visual analog scale (VAS) (83). All subscales differentiated between healthy women (n = 32) and women with RA (n = 44), but after controlling for depression, reduced motivation and mental fatigue no longer differentiated patients from controls (82). Subscales correlated strongly with depression at r = 0.58–0.74 (reduced motivation 0.74) in RA (n = 44) (82). Inflammatory indices (erythrocyte sedimentation rate) were not associated with fatigue subscales in PSS, but in RA, Disease Activity Score scores were moderately associated with general fatigue, physical fatigue, and reduced activity at 0.42–0.47 (n = 49 and 44, respectively) (82). In RA, associations with Short Form 36 (SF-36) pain were stronger for general fatigue, physical fatigue, and reduced activity (−0.51 to −0.61) than for mental fatigue and reduced motivation (−0.23 and −0.40, respectively) (n = 490) (85).

Criterion validity. In AS and RA (n = 812 and 490, respectively), 4 subscales correlated with SF-36 vitality subscale at −0.53 to −0.74, while mental fatigue correlated less strongly (−0.42 and −0.4, respectively), supporting it as a distinct fatigue concept (83,85). Correlations with a fatigue VAS in RA and PSS were strong for general fatigue (0.7 and 0.77, respectively), physical fatigue (0.67 and 0.72, respectively), and reduced activity (0.54 and 0.58, respectively), but moderate for reduced motivation (0.31 and 0.53, respectively) and mental fatigue (0.34 and 0.39, respectively; n = 48 and 490, respectively) (84,85). In FMS, correlations with a fatigue VAS were 0.62 for general fatigue, but 0.32–0.36 for the remaining subscales (n = 165) (87).

Ability to detect change. Three studies report effect sizes (mean change/SD at baseline). In 40 patients with AS randomized to spa therapy, effect sizes were general fatigue 0.82, physical fatigue 0.81, reduced activity 0.28, reduced motivation 0.52, and mental fatigue 0.38, compared to 0.89 in a fatigue VAS (83). In FMS (n = 1,196), a significant improvement was seen in a 20-item–totaled MFI score after milnacipran (88). Also using the 20-item–totaled MFI score (not recommended by the developers), and based on linear regression analysis on comparative fatigue ratings from patients after paired interviews, the effect size required for an average patient to move to a different fatigue category (i.e., much, somewhat or a little, less or more fatigued) is calculated as 0.76 in RA (n = 61) (52). In SLE, again using the totaled 20 items with a range of 20–100, the effect size was 0.59 (95% confidence interval [95% CI] 0.42–0.72), which the authors also present as MFI minimum clinically important difference score of 11.5 (95% CI 8.0–15.0) (53).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The MFI provides a profile of 5 domains of fatigue, and has been used in many long-term and rheumatologic conditions. Internal consistency and test-retest show a range of results, while construct and criterion validity are good. Sensitivity to change was good for general and physical fatigue.

Caveats and cautions. A proportion of patients with cancer had minimum or maximum scores, suggesting there may potentially be significant ceiling and floor effects. Criterion validity was variable across subscales. In rheumatology, the wording of some items may be interpreted as relating to disability or disease activity, and sensitivity to change was weak for some subscales.

Clinical usability. An easy scale to complete in clinic, giving information about fatigue profiles.

Research usability. An easy scale to include in an outcome package. However, potential floor/ceiling effects, and interpretation of some phraseology as relating to broader RA issues rather than fatigue, should be considered.

PEDIATRIC QUALITY OF LIFE (PEDSQL) MULTI-DIMENSIONAL FATIGUE SCALE

Description

Purpose. The PedsQL was developed to measure child and parent perceptions of fatigue in pediatric patients and was published in 2002 (89). It was developed in patients with cancer but is intended as a generic measure for pediatric patients. Versions are available for young adults (ages 18–25), teenagers (ages 13–18), and children (ages 8–12) using developmentally appropriate language, with mirror versions for their parents. A "smiley-face" response version is available for young children (ages 5–7), with a written version for parents, and a parent version for toddlers (ages 2–4).

Content. Covers domains of general fatigue (e.g., I feel tired), fatigue related to sleep/rest (e.g., I feel tired when I wake up in the morning), and cognitive fatigue (e.g., it is hard for me to keep my attention on things).

Number of items. 18 items, giving a total fatigue score and including 3 subscales, each of 6 items (general fatigue, sleep/rest fatigue, and cognitive fatigue).

Response options. 5 response options from "Never a problem" to "Almost always a problem."

Recall period for items. Acute version 7 days, standard version 1 month.

Endorsements. None found for rheumatologic conditions.

Examples of use. The PedsQL Multi-Dimensional Fatigue Scale is a module from PedsQL Measurement model, a modular approach to measuring pediatric health-related quality of life (90). It has been used in studies of mixed rheumatologic disorders, fibromyalgia syndrome (FMS), and juvenile idiopathic arthritis (JIA) (91–93) as well as patients with cancer, cerebral palsy, obesity, cerebral tumours, chronic pain, and multiple sclerosis.

Practical Application

How to obtain. From the web site at URL: http://www.pedsql.org. The PedsQL is free to use in unfunded/internally funded research, otherwise a scale of charges apply depending on funding source (see web site for details).

Method of administration. Child and/or parent self-report, pen and paper. Questionnaires should be read aloud to any children unable to read them. For children unable to understand their age-appropriate version, the preceding version should be offered, or the parent proxy used. For the young child (age 5–7), read questions aloud and show smiley faces response choice page for them to select responses.

Scoring. Raw scores (0–4) are reverse scored and transformed to 0–100 (i.e., 0 = 100, 1 = 75, 2 = 50, 3 = 25, 4 = 0), so that higher scores reflect better health. All 18 items summed and averaged for a total fatigue score, and the 6 items in each subscale summed and averaged for the 3 subscales' scores (general fatigue, sleep/rest fatigue, and cognitive fatigue), all of which range from 0–100. If >50% of the items missing, the scale cannot be scored. Scoring instructions can be downloaded from developers' web site.

Score interpretation. Scores range from 0–100 with higher scores reflecting less fatigue. In terms of normative data, in 52 healthy children (ages 5–18), mean ± SD total fatigue was 80.49 ± 13.33 compared to 76.68 ± 20.523 in children with a range of rheumatologic conditions (n = 152) and 55.48 ± 21.19 in FMS (n = 29), general fatigue in healthy controls was 85.34 ± 14.95 versus 76.82 ± 23.19 and 48.97 ± 25.14, sleep/rest fatigue in healthy controls was 75 ± 18.76 versus 71.77 ± 24.27 and 52.36 ± 21.08, and cognitive fatigue in healthy controls 81.14 ± 17.43 versus 81.30 ± 22.65 and 65.17 ± 24.30 (92). Thus, according to this measure, children with FMS have greater fatigue than children with other rheumatologic conditions, and both groups are worse than healthy controls (although this did not always reach significance in children with broad rheumatologic conditions).

Respondent burden. Estimated at <5 minutes to complete. Items appear easy to read in age-appropriate versions, having undergone cognitive testing.

Administrative burden. Detailed administration instructions need to be read first and suggest some training or practice is required (administration and scoring instructions available on developers' web site). Time to score is not reported, but likely to be ~4–5 minutes to reverse score, transform, sum, and average.

Translations/adaptations. Available in 25 languages (see web site).

Psychometric Information

Method of development. Items and subscales were generated through literature review of adult and pediatric cancer fatigue, patient and parent focus groups, and individual interviews, followed by cognitive interviewing, pretesting, and field testing in cancer (89). Factor analysis appears to have been performed later and data are available for the young adult version (432 university students), where general fatigue and cognitive fatigue loaded on factors 1 and 2, but subscale sleep/rest fatigue loaded across both factors 2 and 3 (94).

Acceptability. In rheumatology, missing item rates of 0.4% and 0.53% are reported for children and 0.7% and 0.8% for parents (91,92). One item might be interpreted in relation to disability or disease activity from rheumatologic conditions rather than fatigue ("I spend a lot of time in bed") and may not be sensitive to rheumatoid arthritis fatigue. Data on floor/ceiling effects could not be located.

Reliability. *Internal consistency.* In rheumatology, Cronbach's alpha ranged from 0.88–0.95 for the total scale and 3 subscales for all age-appropriate versions (n = 163) (91); in FMS, Cronbach's alpha ranged from 0.76–0.94 (n = 29) (92).

Test–retest. No data could be found for the PedsQL.

Interrater reliability. Child and parent (proxy) fatigue scores correlated in a rheumatology population (n = 163) at 0.85–0.93 for total fatigue and all subscales (91).

Validity. *Content validity.* Items were generated through literature review, and patient and parent focus groups and individual interviews in cancer populations (89).

Construct validity. In 175 children with a range of rheumatologic conditions, total fatigue and the 3 subscales correlated strongly with quality of life, pain, physical and psychosocial health, and emotional, social, and school functioning at 0.53–0.91, while all scales had slightly lower, but still positive, associations with daily activities (0.48–0.58); the highest association for total fatigue was with psychological health (0.84), for general fatigue was physical health (0.80), for sleep/rest fatigue was psychological health (0.73), and for cognitive fatigue was poor school functioning (0.77) (91). In 29 children with FMS, the highest correlation for total fatigue and the subscales was always with quality of life (0.69–0.81) (92). Correlation with physician global opinion of disease activity was moderate (−0.30 to −0.39) in a broad rheumatology population (91). Children with inactive JIA (n = 29) showed less fatigue on all subscales than children with active disease (n = 18) (93).

Criterion validity. No data could be found in rheumatology populations. In 432 university students, PedsQL young adult version correlated moderately to strongly with the single item Short Form 8 vitality subscale at 0.56 (general fatigue), 0.54 (total fatigue), 0.4 (cognitive fatigue) and 0.36 (sleep/rest fatigue) (94). In Chinese pediatric patients with cancer, correlation with the fatigue scale-children was −0.45 to −0.61 (n = 108) (95).

Ability to detect change. No sensitivity to intervention data found for any population.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PedsQL is a module from a well-established quality of life measurement system. It reports total fatigue and a range of subscales and has been evaluated in many pediatric long-term conditions. In rheumatology, internal consistency is good, and cognitive fatigue correlates with poor school functioning.

Caveats and cautions. Criterion validity data could not be found for rheumatologic populations, while stability and sensitivity data could not be found for any population. Three subscales appeared to have been generated through a literature review (89) but on later factor analysis (94), the sleep/rest subscale loaded equally across 2 factors, not on a single factor. Total fatigue correlates strongly with psychological status in rheumatology.

Clinical usability. Appears to be a useful tool for clinical use, which is quick to complete.

Research usability. A relatively easy tool to use, but criterion, stability, and sensitivity data are required.

PROFILE OF FATIGUE (PROF)

Description

Purpose. The ProF was developed to characterize patterns of fatigue associated with primary Sjögren's syndrome (PSS) and published in 2003 (96).

Content. Contains somatic fatigue items for needing to rest (e.g., feeling exhausted), difficulty getting started (e.g., hard to get going), low stamina (e.g., hard to keep going), and weak muscles (e.g., feeling weak), and contains mental fatigue items for concentration (e.g., not thinking clearly) and memory (e.g., forgetting things).

Number of items. 16 items giving a total fatigue score, including 6 facets: need rest (4 items), poor starting (4 items), low stamina (2 items), and weak muscles (2 items) can be combined to form the somatic domain. Facets poor concentration (2 items) and poor memory (2) can be combined to form the mental domain.

Response options. 8 response options asking about how patients felt when they were at their worst, ranging from "Not at all" to "As bad as imaginable" (0–7).

Recall period for items. Last 2 weeks.

Endorsements. Developed by the UK Sjögren's Interest Group (96).

Examples of use. Used in PSS, systemic lupus erythematosus (SLE), and rheumatoid arthritis (RA) studies (63,86,96–102).

Practical Application

How to obtain. Obtained from the developers by e-mail: Simon.Bowman@uhb.nhs.uk. The ProF is free to use.

Method of administration. Patient self-report, pen and paper.

Scoring. The 6 facet scores can be reported alone, or combined to form 2 domain scores, or a total score. Facet scores (0–7) are formed by summing and averaging the items in each facet: need rest (items 1–4), poor starting (items 5–8), low stamina (items 9 and 10), weak muscles (11 and 12), poor concentration (13 and 14), and poor memory (15 and 16). Domain scores (0–7) are formed by summing and averaging items 1–12 for somatic fatigue, and items 13–16 for mental fatigue. A total fatigue score (0–7) is created by summing and averaging all 16 items.

Score interpretation. Scores for facets, domains, and total score all range from 0–7 with higher scores reflecting greater fatigue severity. In terms of normative data, in the somatic fatigue domain, the 4 facets had mean values of 1.4–2.2 in 103 healthy controls, and all were significantly different to patients with PSS, RA, and SLE who had mean scores of 2.7–4.4 (n = 18, 18, and 11, respectively); in the mental fatigue domain, the 2 facets both had mean values of 1.5 in healthy controls, which were significantly different to patients with PSS and SLE (mean 2.3–2.5) but not patients with RA (mean 1.9–2.1) (96). The developers used the difference between controls and patients to identify cut points for a "case" of fatigue; a "case" for a facet is someone who scores >2 out of 7 in that facet except for the need rest facet, where ≥3 out of 7 is required (96). A fatigue case for the somatic fatigue domain is a patient who is a "case" in at least 2 of the 4 related facets, while a fatigue case for the mental fatigue domain is a patient who is a "case" in at least 1 of the 2 related facets (96).

Respondent burden. Time to complete not reported but likely to be 4–5 minutes. Items were developed with patients and do not appear difficult, with the possible exception of one ("It's a battle"), which might not be clear to interpret.

Administrative burden. Time to score not reported, likely to be 3–4 minutes to calculate facet, domain, and total scores.

Translations/adaptations. Translated into Swedish using appropriate linguistic methodology (98). A shorter, 6-item ProF was published in 2009 and contains 1 item for each of the 6 facets (97). However, the long version is the most commonly used. A state version ("Right now" rather than "Over the past 2 weeks") has been used (86).

Psychometric Information

Method of development. The ProF contains 16 items from the 64-item Profile of Fatigue and Discomfort–Sicca Symptoms Inventory (96). Draft items were generated using the words of patients with PSS, collected in diaries that were later discussed in focus groups, then subsequently piloted with patients with PSS, RA, and SLE (n = 18, 18, and 11, respectively) who generated 5 clusters of similar statements concerning 4 somatic and 1 mental facet of fatigue (96). The mental component was then split into 2 facets, and all 6 were evaluated in patients with PSS, RA, and SLE (n = 137, 174, and 66, respectively) and controls (n = 103) (96). Factor structure in 82 patients with PSS or RA showed 5 rather than 6 clear factors, with facets need rest and low stamina not being well differentiated, although the somatic and mental fatigue domains were well differentiated (86). The 6 items of the short ProF load on 2 factors, somatic and mental fatigue (97).

Acceptability. Items were developed with patients with PSS and appear easy to read, with missing data reported as only 0.6% (96). The item "It's a battle" might not be answered specifically about fatigue. Authors of one study reported that no floor or ceiling effects were observed (98).

Reliability. *Internal consistency.* Cronbach's alpha for the total fatigue score was 0.97, ranged from 0.91–0.93 for the 2 domains, and from 0.9–0.97 for the 6 facets in patients with PSS (98).

Test–retest. In patients with PSS (n = 12), over a median 3 days (range 0–7 days), the weighted kappa coefficient for total fatigue was 0.63 (interquartile range [IQR] 0.48–0.75); over a median 12 days (range 0–71 days) it was 0.51 (IQR 0.48–0.55; n = 37) (98).

Validity. *Content validity.* The ProF was derived from focus group discussion of PSS patient diaries (96). In the Swedish translation, 19 of 20 patients and both rheumatologists considered the items covered adequate content for PSS fatigue (98).

Construct validity. In terms of somatic fatigue facets, in PSS (n = 18) these correlated with the World Health Organization Quality of Life (WHOQoL) physical domain (−0.62 to −0.69), weak muscles and low stamina correlated with WHOQoL energy at −0.6, and needs rest and poor starting correlated with anxiety and depression at −0.36 to −0.52 (96). In terms of mental fatigue facets, in PSS (n = 18) these correlated with Short Form 36 (SF-36) mental health domain (−0.27 to −0.44), WHOQoL psychological domain (−0.32 to −0.47), and anxiety and depression (−0.34 to −0.48) (96). Sensitivity of facets to classify PSS correctly ranged from 67% (poor memory) to 88% (low stamina), with specificity from 66% (poor starting) to 73% (poor concentration; n = 18) (96). Somatic and mental domains had generally weak associations with accumulative systemic damage in PSS (0.02–0.34) (99). ProF demonstrated that somatic and mental fatigue deteriorate during the day (39 women with PSS or RA) (101).

Criterion validity. In patients with PSS, somatic and mental fatigue domains correlated strongly with SF-36 vitality subscale −0.84 and −0.63, respectively), and with a fatigue visual analog scale (VAS; 0.73 and 0.64, respectively; n = 50) (98). In 82 patients with PSS or RA, all ProF facets correlated strongly with relevant domains from the Multi-Dimensional Fatigue Inventory (0.65–0.86) (86). Patients with PSS, classified as fatigued by a score of >4 on the Fatigue Severity Scale (FSS), had significantly higher mean ± SD ProF scores than those not classified as fatigued on the FSS: somatic fatigue 4.1 ± 1.5 compared to 2.2 ± 1.6, mental fatigue 3.3 ± 1.7 compared to 0.9 ± 1.6 (n = 94) (102). The 6 items of the short ProF correlated with the original 16-item long version domains (0.78 to >0.9) and the short ProF somatic and mental fatigue domains correlated with a fatigue VAS (0.77 and 0.55, respectively; n = 43 PSS) (97).

Ability to detect change. In 17 patients with PSS, somatic fatigue improved significantly in patients randomized to rituximab ($P = 0.009$) but not to placebo ($P = 0.087$; actual data not provided) (100).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The ProF was developed specifically in and for patients with PSS and measures a range of fatigue concepts. Internal consistency is strong and construct and criterion validity are good.

Caveats and cautions. Test–retest reliability is only moderate, and actual data on sensitivity would be helpful. Data on factor structure support 5 rather than 6 factors (facets), and some studies contain relatively few numbers on which to evaluate 6 facets.

Clinical usability. Appears to be a useful, easy tool for clinic use.

Research usability. Appears appropriate for research use, but the caveats above should be considered. There appears to be stronger evidence for the domain structure (somatic, mental fatigue) than the 6-facet structure.

SHORT FORM 36 VITALITY SUBSCALE (SF-36 VT)

The SF-36 is a multidimensional, general health status patient-reported outcome measure (PROM) containing subscales for 8 domains. The detailed review of the entire instrument is presented in the article “Adult Measures of General Health and Health-Related Quality of Life” elsewhere in this issue. This section reports only additional data specific to the SF-36 VT.

Description

Purpose. The SF-36 VT was developed to measure vitality, conceptualized as a single continuum from energy to fatigue, in general and clinical populations, and the complete SF-36 was first published in 1992 (103). The second version was published in 2000 (SF-36v2, see article on Adult Measures of General Health), and in SF-36v2, 1 vitality question has been reworded (from “full of pep” to “full of life”). The SF-12v2, a shorter version published at the same time, also includes a vitality subscale. Most articles do not state whether they have used SF-36 VT or the reworded SF-36v2 VT.

Content. The SF-36 VT covers energy (e.g., feeling full of pep) and fatigue (e.g., feeling worn out), while SF-12 VT contains 1 item on energy.

Number of items. Original and revised versions have 4 items in the SF-36 VT (2 on energy and 2 on fatigue) to produce a single score; SF-12 VT has 1 item (energy).

Response options. In the original SF-36, the vitality subscale had 6 response options ranging from “All of the time” to “None of the time.” In SF-36v2 and SF-12v2, these have been reduced to 5 options to improve psychometric performance (see developer’s web site at URL: http://www.sf-36.org/tools/sf36.shtml).

Recall period for items. 4 weeks, plus a 1-week acute version.

Endorsements. None found for rheumatologic conditions.

Examples of use. SF-36 VT can be aggregated with other subscales to form the mental component score and in earlier literature, SF-36 VT data were not always reported separately. However, with the recent evidence that fatigue is a rheumatology patient priority and part of core data in several conditions (5,6), SF-36 VT data are increasingly being provided. SF-36 VT reports in musculoskeletal studies include data from rheumatoid arthritis (RA), psoriatic arthritis (PsA), ankylosing spondylitis (AS), primary Sjögren’s syndrome (PSS), systemic lupus erythematosus (SLE), fibromyalgia syndrome (FMS), and osteoarthritis (OA) (14,26,52,53,57,70,83,98,104–117). Only a few studies using the SF-12 could be found (all OA), and these did not report the single vitality item separately. Overall, the SF-36 has been used in 14,000 articles, and the revised SF-36v2 in 260, as reported on the developer’s web site at the following URL: http://www.qualitymetric.com/WhatWeDo/GenericHealthSurveys/tabid/184/Default.aspx.

Practical Application

How to obtain. See article on Adult Measures of General Health for web site information on access and cost.

Method of administration. Patient self-report. A range of administration modalities is described in the article on Adult Measures of General Health.

Scoring. As energy items are positive and fatigue items are negative, some items need to be recoded before scoring, then they are summed and transformed to a 0–100 scale (see article on Adult Measures of General Health for details of computerized scoring systems, norm-based algorithms, and handling missing data). The only difference in scoring between the original SF-36 and the SF-36v2 is in the contribution of vitality to the mental and physical component scores. In the original scoring system, the SF-36 VT subscale only contributes to the mental component score but factor analysis in RA (n = 1,030) suggests that vitality correlates equally with both the mental and physical components (0.61 and 0.53, respectively) (113). Therefore, in the revised SF-36v2 scoring system, vitality is now included in both the physical and the mental component scores, but still contributes a larger weighting to the mental component score.

Score interpretation. Scores range from 0–100 with higher scores representing less fatigue. In terms of normative data, age- and sex-based norms are available for many countries (see article on Adult Measures of General Health). Rheumatology studies report mean SF-36 VT scores for healthy controls of 57.4 and 62.2 (n = 77–606) (63,105). This compares to SF-36 VT mean ± SD of 43.4 ± 23.4 in RA, 43.0 ± 24 in AS, 38.9 in PSS, 35.9 ± 23.1 in SLE, 27.1 ± 21.1 in FMS, and 25.7 ± 20.1 in PsA, although SDs are wide (n = 152–13,722) (63,106,109). However, 2 studies report higher vitality than healthy controls in RA, and in patients with OA 2–10 years after arthroplasty (107,108).

Respondent burden. SF-36 VT has only 4 items and therefore would take only 1 minute for respondents to complete, but it is administered with the whole SF-36 questionnaire, which may take up to 10 minutes to complete. The format is not difficult to understand.

Administrative burden. Scoring the SF-36 VT is relatively quick, but it is rarely administered in isolation and scoring the whole SF-36 is more complex and takes longer. Computerized systems are available for purchase from Quality Metric (see article on Adult Measures of General Health).

Translations/adaptations. Available in over 120 languages (see article on Adult Measures of General Health).

Psychometric Information

Method of development. A detailed review of the development of the entire SF-36 is found in the Adult Measures of General Health article. The SF-36 VT items were generated from a review of existing instruments, with the aim of including a balance of favorably and unfavorably worded items (103). There was no patient involvement.

Acceptability. Most items appear acceptable to patients and clearly relate to fatigue or energy. However, in the original SF-36, the item "Full of pep" has the potential to cause confusion in countries where it is not a common term, and has been replaced in SF-36v2 by "Full of life." In RA (n = 1,030), 2.3–5.8% of respondents omitted answers in each of the 4 SF-36 VT items (113). In patients with OA 2–10 years after arthroplasty (n = 58), no floor effects were found but problematic ceiling effects (defined as ≥15%) were found, with 18% of respondents recording best possible scores (107). In contrast, no floor or ceiling effects were found in another report of patients with OA up to 5 years after arthroplasty (n = 59–135) (110). In RA (n = 271), item response theory suggests SF-36 VT covers mainly the less severe range of fatigue severity, in comparison to the Multi-Dimensional Assessment of Fatigue (MAF) and Functional Assessment Chronic Illness Therapy (Fatigue) (FACIT-F), which cover a broader range (57).

Reliability. See article on Adult Measures of General Health for detailed reliability review of SF-36.

Internal consistency. For SF-36 VT, in RA (n = 631), Cronbach's alpha was 0.84–0.88 over 3 time points (57), and in OA (n = 62), all SF-36 domains, including SF-36 VT, had a Cronbach's alpha of 0.75–0.94 (107).

Test–retest. In one OA study (n = 62, mean age 58 years, 2–10 years postarthroplasty), 4-week test–retest reliability of the SF-36 VT was r = 0.92 (107); in contrast, another OA study found very poor 1-week stability at r = 0.03 (n = 21, mean age 70 years) (111). In RA (n = 150), intraclass correlation coefficient for SF-36 VT was 0.91 (95% confidence interval [95% CI] 0.86, 0.94) over 2 weeks (114).

Validity. See article on Adult Measures of General Health for detailed validity review of SF-36.

Content validity. SF-36 VT covers both energy and fatigue, but these may not be opposite ends of a single continuum, as feeling energized is a positive health state rather than the absence of fatigue. Thus, while a person who is not fatigued would score 0 out of 100 on a scale containing 4 fatigue items, they would potentially score 50 on the SF-36 VT by answering "no" to both the energy and fatigue items, due to a lack of energy rather than the presence of fatigue. This is further supported by data from an RA study, where the SF-36v2 VT items loaded across 2 separate factors: "Full of life" and "Lot of energy" loaded on a factor with items feeling happy, peaceful, and healthy, while "Feel tired" and "Worn out" loaded on a factor with items feeling down and feeling sad (n = 401) (115).

Construct validity. In RA (n = 86), SF-36 VT correlated strongly with disability (r = 0.56), and weakly to moderately with physician global assessment, patient global assessment, pain, tender joints, and inflammatory markers (−0.27 to −0.37) (112); correlation with anxiety, depression, and helplessness is reported as 0.28–0.50 (n = 229) (14). SF-36 VT discriminated between patients with RA with low versus moderate Disease Activity Score in 28 joints (DAS28) but not moderate versus high DAS28 (n = 200) (114).

Criterion validity. Data on criterion validity in rheumatology populations are varied for SF-36 VT. For example, correlation with the MAF ranges from very strong (0.79) in RA (70), to strong in OA (−0.54) (55) but only moderate in AS (−0.37) (72). Correlation with a fatigue visual analog scale (VAS) ranges from very strong (0.8) in RA (70), to strong in AS (0.64) (83). Correlation with the facet and domain scores of the Profile of Fatigue ranges from very strong (−0.84) to strong (−0.63) (PSS, n = 50) (98) and with the Multi-Dimensional Fatigue Inventory domains from strong (0.73) to only moderate (0.42) (AS, n = 812) (83). In the evaluation of the Bristol RA Fatigue (BRAF) Multi-Dimensional Questionnaire and its 4 subscales, correlations with SF-36 VT were moderate to strong (−0.40 to −0.68) but in every instance these were lower than the strong correlations between BRAF and MAF or FACIT-F (−0.52 to −0.83) (14).

Ability to detect change. See article on Adult Measures of General Health for detailed review of entire SF-36 ability to detect change. In patients with RA (n = 631) receiving 24 weeks of anti–tumor necrosis factor (TNF) therapy, SF-36 VT showed a mean improvement of 5.2 in patients who did not achieve American College of Rheumatology 20% criteria for improvement in disease activity (ACR20; effect size 0.25), compared to 31.4 in those who achieved

ACR70 (effect size 1.52), which were similar to changes demonstrated by the FACIT-F (57). In PsA (n = 313), 24-week treatment with anti-TNF therapy produced a mean ± SD improvement of 12.8 ± 21 compared to 1.7 ± 19.1 in placebo (59). In SLE (n = 93), while the Chalder Fatigue Questionnaire showed significant improvements in fatigue following exercise compared to relaxation or no intervention, the SF-36 VT did not show improvement, but neither did the Fatigue Severity Scale (FSS) or a fatigue VAS (26). In patients with AS (n = 40) randomized to etanercept or placebo, SF-36 VT showed an effect size of 0.54 for treatment at 1 month and 0.69 at 4 months, while the FSS was not responsive at 1 month (effect size 0.15) but showed a similar effect size at 4 months (0.43) (50). In patients with OA of the hip (n = 135) and knee (n=59) receiving total joint replacement, SF-36 VT showed effect sizes of 1.0 and 0.6, respectively, at 6 months (97). In anti-TNF therapy for patients with RA (n = 258), SF-36 VT showed a change of 16%, which was smaller than change in a fatigue VAS (23%), and changes in tender joint count and patient global assessment (24% and 25%, respectively) (117). Based on linear regression analysis on comparative fatigue ratings from patients after paired interviews, the effect size (mean change/SD at baseline) required for an average patient to move to a different fatigue category (i.e., much, somewhat or a little, less or more fatigued) is calculated as 0.67 in RA (n = 61) (39) and 0.44 (95% CI 0.25, 0.60) in SLE (n = 80), which the authors also present as an SF-36 VT minimum clinically important difference score of −10.7 (95% CI −15.5, −5.9) (53).

Critical Appraisal of Overall Value to the Rheumatology Community

See article on Adult Measures of General Health for overview of the entire SF-36.

Strengths. The SF-36 VT has been used across many rheumatologic conditions and in many studies. Internal consistency, construct validity, and sensitivity are good. The SF-36 VT may be useful when wishing to compare fatigue with other conditions and healthy populations.

Caveats and cautions. In rheumatology populations, there are conceptual concerns over the assumption of fatigue and energy as opposite ends of a single continuum, as energy is a positive health state, rather than an absence of fatigue, which is supported by data demonstrating the 2 energy and 2 fatigue items load on 2 separate factors. There are some reports that vitality is higher in OA and RA than in healthy controls, reports of SF-36 VT ceiling effects, and item response theory suggests that SF-36 VT may not capture higher levels of fatigue. While criterion validity is good with the MAF and a VAS in RA, there are a range of correlations with other fatigue PROMs in rheumatology populations, some as low as 0.37. The conflicting data on test–retest performance in rheumatology (ranging from 0.03–0.92) are concerning.

Clinical usability. The SF-36 VT would be easy to use in clinical practice, but although it was designed for both clinical practice and population surveys, it is not commonly used in clinical care.

Research usability. The SF-36 VT is frequently used in rheumatology research, and provides a global fatigue score. However, the above caveats from data in rheumatology populations should be noted. If the entire SF-36 is being administered in order to capture many health domains to compare with other populations, then researchers may wish to consider whether an additional brief fatigue measure would be helpful.

VISUAL ANALOG SCALES (VAS)

Description

Purpose. Fatigue VAS are unidimensional measures aiming to capture an aspect of fatigue, typically severity or intensity.

Content. Fatigue VAS typically comprise a 100-mm horizontal line, anchored by 2 statements representing extreme ends of a single fatigue continuum (e.g., severity or intensity). However, there is no standardized fatigue VAS for use in rheumatology populations. A systematic review (1996–2004) identified 26 rheumatology studies reporting a fatigue VAS, of which only 4 provided a validation reference, and these related to pain VAS validation; only 10 of 26 were described in detail and only 3 of 26 were identical for content (7). The more recent rheumatology literature, explored for this review, shows the situation continues with multiple VAS versions frequently not described in detail or referenced. Therefore, it appears that researchers often create their own fatigue VAS (stem question and anchors) for individual studies. In terms of content, the stem question may describe tiredness, fatigue, fatigue/tiredness, or unusual fatigue (118–121).

Number of items. A single-item scale.

Response options. Respondents are typically instructed to make a mark across or on the VAS line to describe the point between the 2 anchors that best reflects their fatigue status. Response options are not standardized and depend on the nature of the question, with researchers creating their own. Examples include "Not at all tired" to "Very tired," "No fatigue" to "Total exhaustion," "None" to "As bad as it could be," "No problem" to "Major problem," "Absence of fatigue" to "Worst condition imaginable," "No fatigue" to "Complete fatigue," and "No fatigue" to "Intolerable fatigue" (28,118–120,122–124).

Recall period. Usually 1 week (but often not reported in papers).

Endorsements. None found for rheumatologic conditions.

Examples of use. Used extensively in rheumatologic conditions, e.g., rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), ankylosing spondylitis (AS), psoriatic arthritis (PsA), primary Sjögren's syndrome (PSS), fibromyalgia syndrome (FMS), and osteoarthritis (OA) (14–16,28,70,84,85,88,118–132).

Practical Application

How to obtain. Researchers often create their own VAS.

Method of administration. Self-report by patient using pen and paper.

Scoring. A ruler is used to measure the distance from the left hand anchor to the respondent's mark on the VAS line. While most fatigue VAS range from 0–100 mm, some use a 0–10-cm scale. One variation uses a 15-cm VAS and calculates a score ranging from 0–3, although the rationale for this variation is not provided (125,126). Caution should be taken when scoring VAS, as photocopying can distort (lengthen) the line (133).

Score interpretation. Typically, 0–100 or 0–10 with a higher score representing a greater severity or intensity of fatigue. In terms of normative data, VAS fatigue mean ± SD scores (mm) have been reported in healthy controls (n = 144) as 20.5 ± 0.02 (124). In comparison, examples of rheumatology population means ± SD are 49.7 ± 2.0 in RA, 43.3 ± 2.0 in hand OA, 50.4 ± 30.6 in SLE, 40.8 ± 31.7 in PsA, 74.4 ± 12.9 in PSS, 6.7 ± 2.0 on a scale of 0–10 in AS, and 7.21 ± 1.91 on a scale of 0–10 in FMS (n = 20–202) (84,124,127–130). In 1 RA study, researchers defined fatigue as clinically relevant at VAS ≥20 mm and high fatigue at VAS ≥50 mm (9). In an AS study, researchers defined fatigue as a major symptom at ≥50 mm (83); elsewhere, researchers have defined substantial fatigue in patients with RA, OA, and FMS as ≥2 of 3 on a VAS scaled from 0–3, with a cut off of ≥1 for mild fatigue (125). However, none of these studies report the rationale for the cut points.

Respondent burden. A VAS scale usually takes <1 minute to complete.

Administrative burden. VAS scales are easy to administer and to score. The availability to the patient of their prior VAS score may affect subsequent responses; therefore, researchers should be consistent in whether or not these are available during completion (133).

Translations/adaptations. There is no standard fatigue VAS to translate, but researchers will create their own versions in their own languages. Ideally these should be grounded in patients' words and concepts (13).

Psychometric Information

Method of development. VAS scales have their theoretical foundations in psychological theories of response to sensory stimuli and have a long history in psychometric research to measure subjective states (134). Reports of how the stem question and wording are developed are rare. The Bristol RA Fatigue VAS (3 single items on each of severity, coping, and effects) were developed in collaboration with patients, based on qualitative interviews, then focus groups and cognitive interviewing to design the VAS (13,14): patients chose severity anchors of "No fatigue" to "Totally exhausted," effect anchors of "No effect" to "A great deal of effect," and coping anchors of "Not at all well" to "Very well." This wording is also used in the 3 Bristol Rheumatoid Arthritis Fatigue Numerical Rating Scales (BRAF NRS), which the developers recommend in preference to using VAS, as the NRS versions show stronger psychometric properties and better practical and conceptual considerations than VAS (see BRAF NRS section) (14–16).

Acceptability. In general, most patients find VAS scales easy to understand, and 1 FMS study reports a 99.4% completion rate (87). However, some patients do not understand the VAS measurement concept and may mark above the line or beyond the anchors (133). In RA (n = 7,760, 307), 6.4–9% scored best possible score and 1.8–2% scored worst possible (9,70). Fatigue VAS covers most of the full range of fatigue levels (70).

Reliability. *Test–retest.* In RA over 1–2 days, the intraclass correlation coefficient (ICC) of a fatigue VAS was 0.74 (95% confidence interval [95% CI] 0.65, 0.81; n = 122) (121). In PSS over a median 14 days the ICC was 0.66 (95% CI 0.39, 0.83; n = 48) (84).

Validity. *Content validity.* VAS are unidimensional measures and as they are not standardized, the content largely depends on the construct the researchers wish to explore, and the language they use to capture it.

Construct validity. In a study of 2 RA populations (n = 238 and 274, respectively), fatigue VAS was positively associated with Disease Activity Score at r = 0.43 and r = 0.69, and with pain at r = 0.63 and r = 0.68 (9); in another RA study (n = 22) fatigue VAS correlated very strongly with pain (0.8) and strongly with sleep (0.6) (119). In AS (n = 639), fatigue VAS correlated strongly with axial pain (0.58) but weakly with global pain (0.24) or not with C-reactive protein (−0.07) (120). In FMS (n = 50), fatigue VAS correlated strongly with pain (0.6) but moderately with sleep (0.3), which was not statistically significant (119).

Criterion validity. In RA, fatigue VAS very strongly correlated with the Multi-Dimensional Assessment of Fatigue (MAF) at 0.80, and strongly with Short Form 36 vitality subscale (SF-36 VT) at 0.71 (n = 7,760) (70). In FMS and in PSS, fatigue VAS correlated strongly with Multi-Dimensional Fatigue Inventory (MFI) total (general fatigue) at 0.62 and 0.70, but moderately with MFI mental fatigue and reduced motivation (0.32–0.39), and ranged between moderate and strong for physical fatigue and reduced activity (0.36–0.67) (84,87). In AS (n = 812), fatigue VAS correlated with SF-36 VT at −0.64 (83).

Ability to detect change. In RA (n = 5,155), fatigue VAS was more sensitive to changes in pain and patient global opinion over 6 months than MAF or SF-36 VT, although there was no difference in performance between them in relation to disability or quality of life (70). In anti–tumor necrosis factor (anti-TNF) therapy for patients with RA (n = 391), fatigue VAS showed change of 23% (treatment VAS difference −16.8; 95% CI −22.8, −10.8), similar to improvement in tender joint count and patient global assessment (24% and 25%, respectively) and greater than change in SF-36 VT (16%) (117). Similar changes were seen with anti-TNF therapy in patients with RA and in patients with PsA (n = 30 and 146, respectively) with improvements in fatigue VAS of −17 and −12.0 (9,131). In FMS (n = 40), fatigue VAS (scale 0–10) improved by 2.7 (SD 0.75) after an aquatic exercise program compared to 0.26 (SD 0.35) in controls, with similar changes seen in SF-36 VT (130). In AS (n = 40), a fatigue VAS showed an effect size of 0.89 from spa therapy (83) while in another study (n = 256) improvement in hemoglobin was associated with improvement in fatigue VAS (129). In RA (n = 307), minimum clinically important difference (MCID) for a fatigue VAS of 0–10 was between −0.82 and −1.12 for improvement and 1.13 and 1.26 for worsening, based on a transition question (122); this is similar to the MCID of 10 in a fatigue VAS of 0–100, found by Wells et al (123). In SLE (n = 202), the MCID for a fatigue VAS (0–100) was −13.9 for improvement and 9.1 for worsening, based on a transition question (127). In PsA (n = 200), smaller MCIDs for a fatigue VAS were found at −8.15 for improvement and 3.63 for worsening, also based on a transition question (128). Patients with lower scores required a larger change in their fatigue VAS to report worsening, and people with higher scores required a larger change to perceive improvement, which might be related to floor/ceiling effects or different interpretations at different points in the VAS (122).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. VAS are one of the most frequently used tools to measure fatigue and have been used for many years, with a number of studies supporting their validity for measuring fatigue. They are quick and simple to administer and score, and minimal in terms of respondent burden. In rheumatology, test–retest is good in RA, but weak in PSS, and construct validity is good. Criterion validity is good with MAF, but weaker with SF-36 VT and MFI, while sensitivity to change is good and may be stronger than SF-36 VT. VAS are suitable for use where a global fatigue assessment only is required.

Caveats and cautions. There are many practical and conceptual concerns with VAS, including: VAS length distorts with photocopying; some patients have difficulty understanding the abstract nature of a VAS; the VAS format cannot be administered online or by phone; patients avoid the extreme ends of a VAS; the precision of a 100-mm line may not be appreciated by respondents who tend to consider responses in multiples of 5–10 mm blocks; and when patients' VAS are plotted against their ordinal fatigue scales of none/mild/moderate/severe, VAS scores show considerable overlap across categories (e.g., VAS ratings of 10 and 100 both appear in "moderate" categories) (18,133,135,136). Lack of a standardized fatigue VAS limits comparisons between studies and makes replication across studies difficult, and validation is largely based on accumulative data on a number of differently phrased VAS. A standardized fatigue VAS format, developed with patients with RA (BRAF VAS) has been tested, and found to be marginally less robust than the identical NRS versions; therefore, in view of the other VAS concerns listed here, the NRS versions are recommended by the developers (see BRAF NRS section) (13–16). However, evaluation of the BRAF short scales found that the NRS scored higher than the VAS indicating that the two different patient-reported outcome measures formats are not interchangeable (14). Researchers often create their own VAS, and should take note of studies with pain VAS, which led to recommendations that VAS should be 100-mm long as VAS of <100 mm are inclined to greater error variance, horizontal VAS should be used as they have a more uniform distribution of scores than vertical VAS, anchor wording should be at each end and not below or above the VAS, and that end markers should be placed at right angles to the VAS (not arrows or other markers) (136).

Clinical usability. The fatigue VAS is easy to use in clinical practice, to identify patient concerns and response to treatment. As a single item, VAS are limited in the information that they yield.

Research usability. The fatigue VAS is frequently used in rheumatology research and provides a global fatigue score. However, the above caveats should be noted, and the use of NRS considered. A multidimensional assessment may provide a more complete picture and improve understanding of the clinical relationships of fatigue and hence potential treatment.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

ROLE OF THE STUDY SPONSOR

GlaxoSmithKline had no control over or contribution to the BRAF Scales design, data collection, or analyses. Publication was not contingent on their approval.

REFERENCES

1. Hewlett S, Nicklin J, Treharne GJ. Fatigue in musculoskeletal conditions. Topical Reviews Series 6: No 1. Arthritis Research UK; 2008. URL: http://www.arthritisresearchuk.org/files/6641_25022010173830.pdf.
2. Hewlett S, Cockshott Z, Byron M, Kitchen K, Tipler S, Pope D, et al. Patients' perceptions of fatigue in rheumatoid arthritis: overwhelming, uncontrollable, ignored. Arthritis Rheum 2005;53:697–702.
3. Repping-Wuts H, Uitterhoeve R, van Riel P, van Achterberg T. Fatigue as experienced by patients with rheumatoid arthritis (RA): a qualitative study. Int J Nurs Stud 2008;45:995–1002.
4. Power JD, Badley EM, French MR, Wall AJ, Hawker GA. Fatigue in osteoarthritis: a qualitative study. BMC Musculoskelet Disord 2008; 9:63.
5. Kirwan JR, Minnock P, Adebajo A, Bresnihan B, Choy E, de Wit M, et al. Patient perspective: fatigue as a recommended patient centered outcome measure in rheumatoid arthritis. J Rheumatol 2007;34:174–7.
6. Choy EH, Arnold LM, Clauw DJ, Crofford LJ, Glass JM, Simon LS, et al. Content and criterion validity of the preliminary core dataset for clinical trials in fibromyalgia syndrome. J Rheumatol 2009;36:2330–4.
7. Hewlett S, Hehir M, Kirwan JR. Measuring fatigue in rheumatoid arthritis: a systematic review of scales in use. Arthritis Rheum 2007; 57:429–39.
8. Zonna-Nacach A, Roseman JM, McGwin G Jr, Friedman AW, Baethge BA, Reveille JD, et al. Systemic lupus erythematosus in three ethnic groups. VI. Factors associated with fatigue within 5 years of criteria diagnosis. Lupus 2000;9:101–9.
9. Pollard LC, Choy EH, Gonzalez J, Khoshaba B, Scott DL. Fatigue in rheumatoid arthritis reflects pain, not disease activity. Rheumatology (Oxford) 2006;45:885–9.
10. Hewlett S, Chalder T, Choy E, Cramp F, Davis B, Dures E, et al. Fatigue in rheumatoid arthritis: time for a conceptual model [editorial]. Rheumatology (Oxford) 2011;50:1004–6.
11. Cohen J. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1988. p. 79–81.
12. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcomes measures for use in clinical trials. Health Technol Assess 1998;2:1–74.
13. Nicklin J, Cramp F, Kirwan J, Urban M, Hewlett S. Collaboration with patients in the design of patient-reported outcome measures: capturing the experience of fatigue in rheumatoid arthritis. Arthritis Care Res 2010;62:1552–8.
14. Nicklin J, Cramp F, Kirwan J, Greenwood R, Urban M, Hewlett S. Measuring fatigue in RA: a cross-sectional study to evaluate the Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire, Visual Analog Scales, and Numerical Rating Scales. Arthritis Care Res 2010;62:1559–68.

15. Nicklin J. The development of scales to measure fatigue in people with rheumatoid arthritis. Bristol (UK): University of the West of England; 2009.
16. Dures E, Hewlett S, Kirwan J, Cramp F, Nicklin J, Greenwood R. Test-retest reliability of the Bristol Rheumatoid Arthritis Fatigue Scales (BRAFs) [abstract]. Rheumatology (Oxford) 2011;50:Suppl 3:109.
17. Hewlett S, Dures E, Kirwan JR, Cramp F, Nicklin J, Almeida C, et al. Sensitivity to change of the Bristol Rheumatoid Arthritis Fatigue Scales [abstract]. Arthritis Rheum 2011;63 Suppl.
18. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. Ann Rheum Dis 1978;37:378–81.
19. Chalder T, Berelowitz G, Pawlikowska T, Watts L, Wessely S, Wright D, et al. Development of a fatigue scale. J Psychosom Res 1993;37: 147–53.
20. Cella M, Chalder T. Measuring fatigue in clinical and community settings. J Psychosom Res 2010;69:17–22.
21. Naschitz JE, Rozenbaum M, Fields MC, Enis S, Manor H, Dreyfuss D, et al. Cardiovascular reactivity in fibromyalgia: evidence for pathogenic heterogeneity. J Rheumatol 2005;32:335–9.
22. Tench CM, McCurdie I, White PD, D'Cruz PD. The prevalence and associations of fatigue in systemic lupus erythematosus. Rheumatology (Oxford) 2000;39:1249–54.
23. Strombeck B, Ekdahl C, Manthorpe R, Jacobsson LT. Physical capacity in women with primary Sjogren's syndrome: a controlled study. Arthritis Rheum 2003;49:681–8.
24. Tench C, Bentley D, Vleck V, McCurdie I, White P, D'Cruz D. Aerobic fitness, fatigue, and physical disability in systemic lupus erythematosus. J Rheumatol 2002;29:474–81.
25. White PD, Henderson M, Pearson RM, Coldrick AR, White AG, Kidd BL. Illness behaviour and psychosocial factors in diffuse upper limb pain disorder: a case-control study. J Rheumatol 2003;30:139–45.
26. Tench CM, McCarthy JM, McCurdie I, White PD, D'Cruz PD. Fatigue in systemic lupus erythematosus: a randomized controlled trial of exercise. Rheumatology (Oxford) 2003;42:1050–4.
27. Henderson M, Kidd BL, Pearson RM, White PD. Chronic upper limb pain: an exploration of the biopsychosocial model. J Rheumatol 2005; 32:118–22.
28. Lwin CT, Bishay M, Platts RG, Booth DA, Bowman SJ. The assessment of fatigue in primary Sjogren's syndrome. Scand J Rheumatol 2003; 32:33–7.
29. Vercoulen J, Swanink C, Fennis J, Galama J, van der Meer J, Bleijenberg G. Dimensional assessment of chronic fatigue syndrome. J Psychosom Res 1994;38:383–92.
30. Raterman HG, Hoving JL, Nurmohamed MT, Herenius MM, Sluiter JK, Lems WF, et al. Work ability: a new outcome measure in rheumatoid arthritis? Scand J Rheumatol 2010;39:127–31.
31. Van Hoogmoed D, Fransen J, Bleijenberg G, van Riel P. Physical and psychosocial correlates of severe fatigue in rheumatoid arthritis. Rheumatology (Oxford) 2010;49:1294–302.
32. De Croon EM, Sluiter JK, Nijssen TF, Kammeijer M, Dijkmans BA, Lankhorst GJ, et al. Work ability of Dutch employees with rheumatoid arthritis. Scand J Rheumatol 2005;43:277–83.
33. Hoving JL, Bartelds GM, Sluiter JK, Sadiraj K, Groot I, Lems WF, et al. Perceived work ability, quality of life and fatigue in patients with rheumatoid arthritis after a 6 month course of TNF inhibitors: prospective intervention study and partial economic evaluation. Scand J Rheumatol 2009;38:246–50.
34. Evers AW, Kraaimaat FW, van Riel PL, de Jong AJ. Tailored cognitive-behavioral therapy in early rheumatoid arthritis for patients at risk: an RCT. Pain 2002;100:141–53.
35. Zijlstra TR, Taal E, van de Laar MA, Rasker JJ. Validation of a Dutch translation of the fibromyalgia impact questionnaire. Rheumatology (Oxford) 2007;46:131–4.
36. Van Koulil S, Kraaimaat FW, van Lankveld W, van Riel PL, Evers AW. A patient's persepctive on multidisciplinary treatment gain for fibromyalgia: an indicator for pre-post treatment effects? Arthritis Rheum 2009;61:1626–32.
37. Van Hoogmoed D, Fransen J, Bleijenberg G, van Riel PL. How to assess fatigue in rheumatoid arthritis: validity and reliability of the Checklist Individual Strength [abstract]. Arthritis Rheum 2008;58 Suppl:S868.
38. Knoop H, Van der Meer JW, Bleijenberg G. Guided self-instructions for people with chronic fatigue syndrome: randomized controlled trial. Br J Psychiatry 2008;193:340–1.
39. Beurskens A, Bultmann U, Kant I, Vercoulen J, Bleijenberg G, Swaen G. Fatigue among working people: validity of a questionnaire measure. Occup Environ Med 2000;57:353–7.
40. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The Fatigue Severity Scale: application to patients with multiple sclerosis and systemic lupus erythematosis. Arch Neurol 1989;46:1121–3.
41. Ad Hoc Committee on Systemic Lupus Erythematosus Response Criteria for Fatigue. Measurement of fatigue in systemic lupus erythematosus: a systematic review. Arthritis Rheum 2007;57:1348–57.
42. Schentag CT, Beaton M, Rahman P, Husted J, Gladman DD. Prevalence and correlates of fatigue in psoriatic arthritis (PsA) [abstract]. Arthritis Rheum 2000;43 Suppl:S105.
43. Mancuso CA, Rincon M, Sayles W, Paget SA. Psychosocial variables and fatigue: a longitudinal study comparing individuals with rheumatoid arthritis and healthy controls. J Rheumatol 2006;33:1496–502.
44. Harboe E, Greve OJ, Beyer M, Goransson LG, Tjensvoll AB, Maroni S, et al. Fatigue is associated with cerebral white matter hyperintensities in patients with systemic lupus erythematosus. J Neurol Neurosurg Psychiatry 2008;79:199–201.
45. Schentag CT, Cichon J, MacKinnon A, Gladman DD, Urowitz MB. Validation and normative data for the 0–10 point scale version of the fatigue severity scale (FSS) [abstract]. Arthritis Rheum 2000;43 Suppl: S177.
46. Mattsson M, Moller B, Lundberg IE, Gard G, Bostrom C. Reliability and validity of the Fatigue Severity Scale in Swedish for patients with systemic lupus erythematosus. Scand J Rheumatol 2008;37:269–77.
47. Austin JS, Maisiak RS, Macrina DM, Heck LW. Health outcome improvements in patients with systemic lupus erythematosus using two telephone counseling interventions. Arthritis Care Res 1996;9:391–9.
48. Omdal R, Mellgren SI, Koldingsnes W, Jacobsen EA, Husby G. Fatigue in patients with systemic lupus erythematosus: lack of associations to serum cytokines, anti-phospholipid antibodies or other disease characteristics. J Rheumatol 2002;29:482–6.
49. Chandran V, Behella S, Schentag C, Gladman DD. Functional Assessment of Chronic Illness Therapy-Fatigue Scale is valid in patients with psoriatic arthritis. Ann Rheum Dis 2007;66:936–9.
50. Wanders AJ, Gorman JD, Davis JC, Landewe RB, van der Heijde DM. Responsiveness and discriminative capacity of the assessments in ankylosing spondylitis disease-controlling antirheumatic therapy core set and other outcome measures in a trial of etanercept in ankylosing spondylitis. Arthritis Rheum 2004;51:1–8.
51. Ramsey-Goldman R, Schilling EM, Dunlop D, Langman C, Greenland P, Thomas RJ, et al. A pilot study on the effects of exercise in patients with systemic lupus erythematosus. Arthritis Care Res 2000;13:262–9.
52. Pouchot J, Kherani RB, Brant R, Lacaille D, Lehman AJ, Ensworth S, et al. Determination of the minimal clinically important difference for seven fatigue measures in rheumatoid arthritis. J Clin Epidemiol 2008; 61:705–13.
53. Goligher EC, Pouchot J, Brant R, Kherani R, Avina-Zubieta JA, Lacaille D, et al. Minimal clinically important difference for seven fatigue measures in patients with systemic lupus erythematosus. J Rheumatol 2008;35:635–42.
54. Cross M, Lapsley H, Barcenilla A, Brooks P, March L. Association between measures of fatigue and health-related quality of life in rheumatoid arthritis and osteoarthritis. The Patient: Patient-Centered Outcomes Research 2008;1:97–104.
55. Schwartz JE, Jandorf L, Krupp LB. The measurement of fatigue: a new instrument. J Psychosom Res 1993;37:753–62.
56. Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E. Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) Measurement System. J Pain Symptom Manage 1997;13:63–74.
57. Cella D, Yount S, Sorensen M, Chartash E, Sengupta N, Grober J. Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. J Rheumatol 2005;32:811–9.
58. Walker J, Gordon T, Lester S, Downie-Doyle S, McEvoy D, Pile K, et al. Increased severity of lower urinary tract symptoms and daytime somnolence in primary Sjogren's syndrome. J Rheumatol 2003; 30:2406–12.
59. Gladman DD, Mease PJ, Cifaldi MA, Perdok RJ, Sasso E, Medich J. Adalimumab improves joint-related and skin-related functional impairment in patients with psoriatic arthritis: patient-reported outcomes of the Adalimumab effectiveness in psoriatic arthritis trial. Ann Rheum Dis 2007;66:163–8.
60. Mittendorf T, Dietz B, Sterz R, Kupper H, Cifaldi MA, von der Schulenburg JM. Improvement and long-term maintenance of quality of life during treatment with adalimumab in severe rheumatoid arthritis. J Rheumatol 2007;34:2343–50.
61. Cohen SB, Emery P, Greenwald MW, Dougados M, Furie RA, Genovese MC, et al. Rituximab for rheumatoid arthritis refractory to anti–tumor necrosis factor therapy: results of a multicenter, randomized, double-blind, placebo-controlled, phase III trial evaluating primary efficacy and safety at twenty-four weeks. Arthritis Rheum 2006;54: 2793–806.
62. Weinblatt ME, Keystone EC, Furst DE, Moreland LW, Weisman MH, Birbara CA, et al. Adalimumab, a fully human anti–tumor necrosis factor α monocolonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate. Arthritis Rheum 2003;48:35–45.
63. Segal B, Bowman SJ, Fox PC, Vivino FB, Murukutla N, Brodscholl J, et al. Primary Sjogren's syndrome: health experiences and predictors

of health quality among patients in the United States. Health Qual Life Outcomes 2009;7:46.

64. Cella D, Lai JS, Chang CH, Peterman A, Slavin M. Fatigue in cancer patients compared with fatigue in the general United States population. Cancer 2002;94:528–38.
65. Nikolaus S, Bode C, Taal E, van de Laar MA. Four different patterns of fatigue in rheumatoid arthritis patients: results of a Q-sort study. Rheumatology (Oxford) 2010;49:2191–9.
66. Tack B. Dimensions and correlates of fatigue in older adults with rheumatoid arthritis. San Francisco: University of California; 1991.
67. Piper B, Lindsey A, Dodd M, Ferketich S, Paul S, Weller S. The development of an instrument to measure the subjective dimension of fatigue. In: Funk S, Tornquist E, Champagne M, Wiese R, editors. Key aspects of comfort: management of pain, fatigue, and nausea. New York: Springer; 1989. p. 199–207.
68. Belza BL. Comparison of self-reported fatigue in rheumatoid arthritis and controls. J Rheumatol 1995;22:639–43.
69. Belza B, Henke C, Yelin E, Epstein W, Gilliss C. Correlates of fatigue in older adults with rheumatoid arthritis. Nurs Res 1993;42:93–9.
70. Wolfe F. Fatigue assessments in rheumatoid arthritis: comparative performance of visual analog scales and longer fatigue questionnaires in 7,760 patients. J Rheumatol 2004;31:1896–902.
71. Stebbings S, Herbison P, Doyle TC, Treharne G, Highton J. A comparison of fatigue correlates in rheumatoid arthritis and osteoarthritis: disparity in associations with disability, anxiety and sleep disturbance. Rheumatology (Oxford) 2010;49:361–7.
72. Turan Y, Duruoz MT, Bal S, Guvenc A, Cerrahoglu L, Gurgan A. Assessment of fatigue in patients with ankylosing spondylitis. Rheumatol Int 2007;27:847–52.
73. Neuberger GB, Press AN, Lindsley HB, Hinton R, Cagle PE, Carlson K, et al. Effects of exercise on fatigue, aerobic fitness, and disease activity measures in persons with rheumatoid arthritis. Res Nurs Health 1997; 20:195–204.
74. Arnold LM, Zlateva G, Sadosky A, Emir B, Whalen E. Correlations between fibromyalgia symptom and function domains and patient global impression of change: a pooled analysis of three randomized, placebo-controlled trials of Pregabalin. Pain Med 2011;12:260–7.
75. Arnold LM, Chatamra K, Hirsch I, Stoker M. Safety and efficacy of Esreboxetine in patients with fibromyalgia: an 8-week, multicenter, randomized, double-blind, placebo-controlled study. Clin Ther 2010; 32:1618–32.
76. Yacoub YI, Amine B, Laatiris A, Abouqal R, Hajjaj-Hassouni N. Assessment of fatigue in Moroccan patients with ankylosing spondylitis. Clin Rheumatol 2010;29:1295–9.
77. Jump RL, Fifield J, Tennen H, Reisine S, Giuliano AJ. History of affective disorder and the experience of fatigue in rheumatoid arthritis. Arthritis Rheum 2004;51:239–45.
78. Bouzubar FF. Self-reported fatigue in individuals with knee osteoarthritis. Pittsburgh: University of Pittsburgh; 2003.
79. Meek PM, Nail LM, Barsevick A, Schwartz AL, Stephen S, Whitmer K, et al. Psychometric testing of fatigue instruments for use with cancer patients. Nurs Res 2000;49:181–90.
80. Smets E, Garssen B, Bonke B, De Haes J. The Multidimensional Fatigue Inventory (MFI): psychometric qualities of an instrument to assess fatigue. J Psychosomatic Res 1995;39:315–25.
81. Smets E, Garssen B, Cull A, de Haes J. Application of the multidimensional fatigue inventory (MFI-20) in cancer patients receiving radiotherapy. Br J Cancer 1996;73:241–5.
82. Barendregt P, Visser M, Smets E, Tulen J, van den Meiracker A, Boomsma F, et al. Fatigue in primary Sjogren's syndrome. Ann Rheum Dis 1998;57:291–5.
83. Van Tubergen A, Coenen J, Landewe R, Spoorenberg A, Chorus A, Boonen A, et al. Assessment of fatigue in patients with ankylosing spondylitis: a psychometric analysis. Arthritis Rheum 2002;47:8–16.
84. Forsblad d'Elia H, Rehnberg E, Kvist G, Ericsson A, Konttinen YT, Mannerkorpi K. Fatigue and blood pressure in primary Sjogren's syndrome. Scand J Rheumatol 2008;37:284–92.
85. Rupp I, Boshuizen HC, Jacobi CE, Dinant HJ, van den Bos GA. Impact of fatigue on health-related quality of life in rheumatoid arthritis. Arthritis Rheum 2004;51:578–85.
86. Goodchild CE, Treharne GJ, Booth DA, Kitas GD, Bowman SJ. Measuring fatigue among women with Sjogren's syndrome or rheumatoid arthritis: a comparison of the Profile of Fatigue (ProF) and the Multidimensional Fatigue Inventory (MFI). Musculoskelet Care 2008;6: 31–48.
87. Ericsson A, Mannerkorpi K. Assessment of fatigue in patients with fibromyalgia and chronic widespread pain: reliability and validity of the Swedish version of the MFI-20. Disabil Rehabil 2007;29:1665–70.
88. Clauw DJ, Mease P, Palmer RH, Gendreau RM, Wang Y. Milnacipran for the treatment of fibromyalgia in adults: a 15 week, multicenter, randomized, double-blind, placebo-controlled, multiple-dose clinical trial. Clin Ther 2008;30:1988–2004.
89. Varni JW, Burwinkle TM, Katz ER, Meeske K, Dickinson P. The PedsQL in pediatric cancer: reliability and validity of the Pediatric Quality of Life Inventory Generic Core scales, multidimensional fatigue scale, and cancer module. Cancer 2002;94:2090–106.
90. Varni JW, Seid M, Rode CA. The PedsQL: measurement model for the Pediatric Quality of Life Inventory. Med Care 1999;37:126–39.
91. Varni JW, Burwinkle TM, Szer IS. The PedsQL Multidimensional Fatigue Scale in pediatric rheumatology: reliability and validity. J Rheumatol 2004;31:2494–500.
92. Varni JW, Burwinkle TM, Limbers CA, Szer IS. The PedsQL as a patient-reported outcome in children and adolescents with fibromyalgia: an analysis of OMERACT domains. Health Qual Life Outcomes 2007;5:9.
93. Ringold S, Wallace CA, Rivara FP. Health-related quality of life, physical function, fatigue and disease activity in children with established polyarticular juvenile idiopathic arthritis. J Rheumatol 2009; 36:1330–6.
94. Varni JW, Limbers CA. The PedsQL Multidimensional Fatigue Scale in young adults: feasibility, reliability and validity in a University student population. Qual Life Res 2008;17:105–14.
95. Chiang YC, Hinds PS, Yeh CH, Yang CP. Development and psychometric testing of a Chinese version of the Fatigue Scale-Children in Taiwan. J Clin Nursing 2008;17:1201–10.
96. Bowman SJ, Booth DA, Platts RG, and the UK Sjogren's Interest Group. Measurement of fatigue and discomfort in primary Sjorgren's syndrome using a new questionnaire tool. Rheumatology (Oxford) 2004; 43:758–64.
97. Bowman SJ, Hamburger J, Richards A, Barry RJ, Sauz S. Patient-reported outcomes in primary Sjogren's syndrome: comparison of the long and short versions of the Profile of Fatigue and Discomfort-Sicca Symptoms Inventory. Rheumatology (Oxford) 2009;48:140–3.
98. Strombeck B, Theander E, Jacobsson LT. Assessment of fatigue in primary Sjogren's syndrome: the Swedish version of the Profile of Fatigue. Scand J Rheumatol 2005;34:455–9.
99. Barry RJ, Sutcliffe N, Isenberg DA, Price E, Goldblatt F, Adler M, et al. The Sjogren's Syndrome Damage Index: a damage index for use in clinical trials and observational studies in primary Sjogren's syndrome. Rheumatology (Oxford) 2008;47:1193–8.
100. Dass S, Bowman SJ, Vital EM, Ikeda K, Pease CT, Hamburger J, et al. Reduction of fatigue in Sjogren's syndrome with rituximab: results of a randomised, double-blind, placebo-controlled pilot study. Ann Rheum Dis 2008;67:1541–4.
101. Goodchild EC, Treharne GJ, Booth DA, Bowman SJ. Daytime patterning of fatigue and its associations with the previous night's discomfort and poor sleep among women with primary Sjogren's syndrome or rheumatoid arthritis. Musculoskelet Care 2010;8:107–17.
102. Segal B, Thomas W, Rogers T, Leon JM, Hughes P, Patel D, et al. Prevalence, severity, and predictors of fatigue in subjects with primary Sjögren's syndrome. Arthritis Rheum 2008;59:1780–7.
103. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): conceptual framework and item selection. Med Care 1992;30: 473–83.
104. Keystone E, Burmester GR, Furie R, Loveless JE, Emery P, Kremer J, et al. Improvement in patient-reported outcomes in a Rituximab trial in patients with severe rheumatoid arthritis refractory to anti–tumor necrosis factor therapy. Arthritis Rheum 2008;59:785–93.
105. Gilboe IM, Kvien TK, Husby G. Disease course in systemic lupus erythematosus: changes in health status, disease activity and organ damage after 2 years. J Rheumatol 2001;28:266–74.
106. Wolfe F, Michaud K, Li T, Katz RS. EQ-5D and SF-36 quality of life measures in systemic lupus erythematosus: comparisons with rheumatoid arthritis, non-inflammatory rheumatic disorders and fibromyalgia. J Rheumatol 2010;37:296–304.
107. Soderman P, Malchau H. Validity and reliability of the Swedish WOMAC osteoarthritis index: a self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP). Acta Orthop Scand 2000;71:39–46.
108. Tuttleman M, Pillemer SR, Tilley BC, Fowler SE, Budkley LM, Alarcon GM, et al. A cross sectional assessment of health status instruments in patients with rheumatoid arthritis participating in a clinical trial. J Rheumatol 1997;24:1910–5.
109. Dagfinrud H, Vollestad NK, Loge JH, Kvien TK, Mengshoel AM. Fatigue in patients with ankylosing spondylitis: a comparison with the general population and associations with clinical and self-reported measures. Arthritis Rheum 2005;53:5–11.
110. Busija L, Osborne RH, Nilsdotter A, Buchbinder R, Roos EM. Magnitude and meaningfulness of change in SF-36 scores in four types of orthopaedic surgery. Health Qual Life Outcomes 2008;6:55.
111. Davey RC, Matthes Edwards S, Cochrane T. Test-retest reliability of lower extremity functional and self-reported measures in elderly with osteoarthritis. Adv Physiother 2003;5:155–60.
112. Birrell FN, Hassell AB, Jones PW, Dawes PT. How does the Short Form 36 Health Questionnaire (SF-36) in rheumatoid arthritis (RA)

relate to RA outcome measures and SF-36 population values? A cross-sectional study. Clin Rheumatol 2000;19:195–9.

113. Loge JH, Kaasa S, Hjermstad MJ, Kvien TK. Translation and performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. I. Data quality, scaling assumptions, reliability and construct validity. J Clin Epidemiol 1998;51:1069–76.

114. Linde L, Sorensen J, Ostergaard M, Horslev-Peterson, Hetland ML. Health-related quality of life: validity, reliability and responsiveness of SF-36, EQ-15D, EQ-5D, RAQoL and HAQ in patients with rheumatoid arthritis. J Rheumatol 2008;35:1528–37.

115. Koh ET, Leong KP, Tsou IY, Lim VH, Pong LY, Chong SY, et al. The reliability, validity and sensitivity to change of the Chinese version of SF-36 in oriental patients with rheumatoid arthritis. Rheumatology (Oxford) 2006;45:1023–8.

116. Saad AA, Ashcroft DM, Watson KD, Symmons DP, Noyce PR, Hyrich KL. Improvements in quality of life and functional status in patients with psoriatic arthritis receiving anti–tumor necrosis factor therapies. Arthritis Care Res 2010;62:345–53.

117. Wells G, Li T, Maxwell L, MacLean R, Tugwell P. Responsiveness of patient reported outcomes including fatigue, sleep quality, activity limitation, and quality of life following treatment with abatacept for rheumatoid arthritis. Ann Rheum Dis 2008;67:260–5.

118. Riemsma RP, Rasker JJ, Taal E, Griep EN, Wouters JM, Wiegman O. Fatigue in rheumatoid arthritis: the role of self-efficacy and problematic social support. Br J Rheumatol 1998;37:1042–6.

119. Mengshoel AM, Forre O. Pain and fatigue in patients with rheumatic disorders. Clin Rheumatol 1993;12:515–22.

120. Dernis-Labous E, Messow M, Dougados M. Assessment of fatigue in the management of patients with ankylosing spondylitis. Rheumatology (Oxford) 2003;42:1523–8.

121. Rohekar G, Pope J. Test-retest reliability of patient global assessment and physician global assessment in rheumatoid arthritis. J Rheumatol 2009;36:2178–82.

122. Khanna D, Pope J, Khanna PP, Maloney M, Samedi N, Norrie D, et al. The minimally important difference for the fatigue visual analog scale in patients with rheumatoid arthritis followed in an academic clinical practice. J Rheumatol 2008;35:2339–43.

123. Wells G, Li T, Maxwell L, MacLean R, Tugwell P. Determining the minimal clinically important differences in activity, fatigue, and sleep quality in patients with rheumatoid arthritis. J Rheumatol 2007;34:280–9.

124. Slatkowsky-Christensen B, Mowinckel P, Loge JH, Kvien TK. Health-related quality of life in women with symptomatic hand osteoarthritis: a comparison with rheumatoid arthritis patients, healthy controls, and normative data. Arthritis Rheum 2007;57:1404–9.

125. Wolfe F, Hawley D, Wilson K. The prevalence and meaning of fatigue in rheumatic disease. J Rheumatol 1996;23:1407–17.

126. Wolfe F. Determinants of WOMAC function, pain and stiffness scores: evidence for the role of low back pain, symptoms counts, fatigue and depression in osteoarthritis, rheumatoid arthritis and fibromyalgia. Rheumatology (Oxford) 1999;38:355–61.

127. Colangelo KJ, Pope JE, Peschken C. The minimally important difference for patient reported outcomes in systemic lupus erythematosus including the HAQ-DI, pain, fatigue, and SF-36. J Rheumatol 2009;36:2231–7.

128. Kwok T, Pope JE. Minimally important difference for patient-reported outcomes in psoriatic arthritis: health assessment questionnaire and pain, fatigue, and global visual analog scales. J Rheumatol 2010;37:1024–8.

129. Braun J, van der Heijde D, Doyle MK, Han C, Deodhar A, Inman R, et al. Improvement in hemoglobin levels in patients with ankylosing spondylitis treated with Infliximab. Arthritis Rheum 2009;61:1032–6.

130. Ide MR, Laurindo IM, Rodrigues-Junior AL, Tanaka C. Effect of aquatic respiratory exercise-based program in patients with fibromyalgia. Int J Rheum Dis 2008;11:131–40.

131. Heiberg MS, Kaufmann C, Rodevand E, Mikkelsen K, Koldingsnes W, Mowinckel P, et al. The comparative effectiveness of anti-TNF therapy and methotrexate in patients with psoriatic arthritis: 6 month results from a longitudinal, observational, multicentre study. Ann Rheum Dis 2007;66:1038–42.

132. Heiberg T, Finset A, Uhlig T, Kvien TK. Seven year changes in health status and priorities for improvement of health in patients with rheumatoid arthritis. Ann Rheum Dis 2005;64:191–5.

133. Bellamy N. Musculoskeletal clinical metrology. London: Kluwer Academic Publishers; 1993.

134. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. Health Technol Assess 1999;3:9.

135. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. 4th ed. Oxford: Oxford University Press; 2008.

136. Huskisson EC. Visual analogue scales. In: Melzack R, editor. Pain measurement and assessment. New York: Raven Press; 1983.

Summary Table for Fatigue Measures

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire	Measure severity, impact, and dimensions of fatigue in rheumatoid arthritis	Patient self-report	4–5 minutes	2–3 minutes	Higher = worse	Internal consistency: strong; test–retest: strong	Content validity: strong; construct validity: strong; criterion validity: strong	Good, based on data still under peer review	Developed with rheumatoid arthritis patients, contains multiple subscales	Currently, data only available from developmental article; reliability and sensitivity article awaited
Bristol Rheumatoid Arthritis Fatigue Numerical Rating Scales (severity, effect, and coping)	Measure severity, impact, and coping with fatigue in rheumatoid arthritis	Patient self-report	1 minute	1 minute	Severity and effect: higher = worse; coping higher = better	Severity and effect test–retest: strong; coping test–retest: moderate	Content validity: strong; construct validity: strong; criterion validity: strong, moderate for coping	Good, based on data still under peer review	Developed with rheumatoid arthritis patients, measures severity, effect, and coping	Currently, data only available from developmental article; reliability and sensitivity article awaited
Chalder Fatigue Questionnaire	Measure severity in hospital and community populations	Patient self-report	2–3 minutes	2–3 minutes	Higher = worse	Internal consistency: strong; test–retest: strong in other populations	Content validity: good; construct validity: moderate; criterion validity: moderate	Good	Measures different subscales	No reliability data for rheumatology; does not always differentiate between rheumatology patients and controls
Checklist Individual Strength	Measure aspects of fatigue in chronic fatigue syndrome	Patient self-report	4–5 minutes	4–5 minutes	Higher = worse	Internal consistency: strong; test–retest: strong	Content validity: moderate; construct validity: strong; criterion validity: strong	Good	Evaluated in many long-term conditions, contains multiple subscales	3 items might be confounded by rheumatology disease or disability
Fatigue Severity Scale	Measure disabling fatigue in multiple sclerosis and systemic lupus erythematosus	Patient self-report	2–3 minutes	2–3 minutes	Higher = worse	Internal consistency: strong; test–retest: strong	Content validity: strong; construct validity: strong; criterion validity: strong	Good	Recommended fatigue scale for systemic lupus erythematosus	A 10-point version is being used in psoriatic arthritis
Functional Assessment Chronic Illness Therapy (Fatigue)	Measure fatigue in anemic oncology patients, later tested in chronic illness	Patient self-report	3–4 minutes	3–4 minutes	Higher = better	Internal consistency: strong; test–retest: strong	Content validity: moderate; construct validity: strong; criterion validity: strong	Good	Evaluated in several rheumatologic conditions	Phrasing of 4 of 13 items potentially confounded by rheumatologic conditions
Multi-Dimensional Assessment of Fatigue	Measure multiple dimensions of fatigue in adults with rheumatoid arthritis	Patient self-report	5–8 minutes	4–5 minutes	Higher = worse	Internal consistency: strong; test–retest: strong	Content validity: moderate; construct validity: strong; criterion validity: strong	Good	Rheumatoid arthritis specific but also evaluated in a range of long-term conditions	High levels of missing data reported; items might be answered in relation to disability
Multi-Dimensional Fatigue Inventory	Measure cancer fatigue using a multidimensional, short questionnaire without any somatic items	Patient self-report	4–5 minutes	4–5 minutes	Higher = worse	Internal consistency: strong; test–retest: strong	Content validity: moderate; construct validity: strong; criterion validity: moderate and variable	Good	Contains multiple subscales, evaluated in long-term and rheumatology conditions	Mental fatigue scale does not correlate strongly with other subscales; potential floor and ceiling effects; criterion validity variable across subscales; phrasing may be confounded by rheumatologic conditions
Pediatric Quality of Life Multi-Dimensional Fatigue Scale	Measure child and parent perceptions of fatigue in pediatric patients; developed in cancer, intended as generic	Patient self-report	4–5 minutes	4–5 minutes	Higher = better	Internal consistency: strong; test–retest: strong	Content validity: moderate; construct validity: strong to moderate; criterion validity: moderate in healthy children	No data in any disease could be located	Measures multiple domains; part of a well-established measurement system	No rheumatology criterion data, nor any sensitivity data; correlates very strongly with psychological health in rheumatology
Profile of Fatigue	Characterize patterns of fatigue in primary Sjögren's syndrome	Patient self-report	4–5 minutes	4–5 minutes	Higher = worse	Internal consistency: strong; test–retest: moderate	Content validity: strong; construct validity: moderate; criterion validity: strong	Good	Primary Sjögren's syndrome specific, developed with patients. Contains multiple subscales	Test–retest reliability data moderate; some studies have small numbers to test 6 facets
Short Form 36 vitality subscale	Measure vitality (energy and fatigue) in general and clinical populations	Patient self-report	1 minute	1–2 minutes	Higher = better	Internal consistency: strong; test–retest: variable, very weak to strong	Content validity: moderate; construct validity: strong; criterion validity: variable, moderate to strong	Good	Widely used, can compare across conditions and general population	Concerns over concepts of fatigue vs. energy, criterion validity, and test–retest in rheumatology populations
Visual analog scales	Measure whatever fatigue constructs are required	Patient self-report	1 minute	1 minute	Higher = worse	Test–retest: strong	Content validity: no standard format; construct validity: strong; criterion validity: variable, moderate to strong	Good	Widely used, a quick screening patient-reported outcome	Standardized version developed but was less robust than numerical rating scale version; visual analog scale can be confusing for some patients; photocopying distorts the line

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S420–S430
DOI 10.1002/acr.20637

Measures of General Pediatric Quality of Life

Child Health Questionnaire (CHQ), DISABKIDS Chronic Generic Measure (DCGM), KINDL-R, Pediatric Quality of Life Inventory (PedsQL) 4.0 Generic Core Scales, and Quality of My Life Questionnaire (QoML)

STEPHANIE E. HULLMANN, JAMIE L. RYAN, RACHELLE R. RAMSEY, JOHN M. CHANEY, AND LARRY L. MULLINS

CHILD HEALTH QUESTIONNAIRE (CHQ)

Description

Purpose. To measure health-related quality of life (HRQOL) in children and adolescents ages 5–18 years. This measure consists of child report (ages 10–18 years) and 2 versions of parent-proxy report (ages 5–18 years) of the child's HRQOL. It can be used with healthy children and those with both acute and chronic health conditions.

Content. Assesses for 14 physical and psychosocial domains: general health perceptions, physical functioning, role/social physical functioning, bodily pain, role/social emotional functioning, role/social behavioral functioning, parent impact-time, parent impact-emotional, self-esteem, mental health, behavior, family activities, family cohesion, and change in health.

Number of items. The child-report questionnaire (CHQ-CF87) consists of 87 items. The long parent-report questionnaire (CHQ-PF50) consists of 50 items, and the short parent-report questionnaire (CHQ-PF28) consists of 28 items.

Response options/scale. The response options for the CHQ are ordinal scales that vary by the item. Each item consists of 4–6 response options. Additionally, each scale consists of varying numbers of items.

Recall period for items. Varies by subscale. Most scales have a recall period of 4 weeks. The change in health subscale has a recall period of 1 year, and the global health, general health perception, and family cohesion subscales ask about the child's health "in general."

Endorsements. No information.

Examples of use. Apaz MT, Saad-Magalhaes C, Pistorio A, Ravelli A, de Oliveira Sato J, Marcantoni MB, et al, for the Paediatric Rheumatology International Trials Organisation. Health-related quality of life of patients with juvenile dermatomyositis: results from the Paediatric Rheumatology International Trials Organisation multinational quality of life cohort study. Arthritis Rheum 2009; 61:509–17.

Brunner HI, Higgins GC, Wiers K, Lapidus SK, Olson JC, Onel K, et al. Health-related quality of life and its relationship to patient disease course in childhood-onset systemic lupus erythematosus. J Rheumatol 2009;36: 1536–45 (1).

Gutierrez-Suarez R, Pistorio A, Cespedes Cruz A, Norambuena X, Flato B, Rumba I, et al. Health-related quality of life of patients with juvenile idiopathic arthritis coming from 3 different geographic areas: the PRINTO multinational quality of life cohort study. Rheumatology (Oxford) 2007;46:314–20 (2).

Oliveira S, Ravelli A, Pistorio A, Castell E, Malattia C, Prieur AM, et al, for the Pediatric Rheumatology International Trials Organization (PRINTO). Proxy-reported health-related quality of life of patients with juvenile idiopathic arthritis: the Pediatric Rheumatology International Trials Organization multinational quality of life cohort study. Arthritis Rheum 2007;57:35–43 (3).

Ruperto N, Buratti S, Duarte-Salazar C, Pistorio A, Reiff A, Bernstein B, et al. Health-related quality of life in juvenile-onset systemic lupus erythematosus and its relationship to disease activity and damage. Arthritis Rheum 2004; 51:458–64.

Selvaag AM, Flato B, Lien G, Sorskaar D, Vinje O, Forre O. Measuring health status in early juvenile idiopathic arthritis: determinants and responsiveness of the Child Health Questionnaire. J Rheumatol 2003;30:1602–10 (4).

Takken T, Elst E, Spermon N, Helders PJ, Prakken AB, van der Net J. The physiological and physical determinants of functional ability measures in children with juvenile dermatomyositis. J Rheumatol 2002;42:591–5.

Stephanie E. Hullmann, MS, Jamie L. Ryan, MS, Rachelle R. Ramsey, MS, John M. Chaney, PhD, and Larry L. Mullins, PhD: Oklahoma State University, Stillwater.

Address correspondence to Stephanie E. Hullmann, MS, Oklahoma State University, 116 North Murray Hall, Stillwater, OK 74078. E-mail: stephanie.hullmann@okstate.edu.

Submitted for publication January 23, 2011; accepted in revised form May 10, 2011.

Practical Application

How to obtain. The CHQ scales can be obtained from the authors at www.healthact.com. The licensing fee is based upon the proposed use of the questionnaires, funding source, sample size, number of administrations, number of sites, start and end dates of the project, and the language.

Method of administration. Parents and children (ages 10–18 years) may self-administer the CHQ after instructions from the administrator.

Scoring. Overall means for the individual CHQ scales and items can be derived using a simple summated rating approach. This method yields a profile for each of the 14 health concepts. In addition, the individual scale scores can be aggregated to derive 2 summary component scores: the physical functioning and psychosocial health summary scores. Scores are transformed to a 0–100 scale, with a mean ± SD of 50 ± 10. The CHQ Scoring and Interpretation Manual is available on CD and is required for scoring and interpretation.

Score interpretation. Range on subscales and the overall scale is 0–100, where 0 = worst possible health state and 100 = best possible health state. Individual or population means of parent-reported quality of life can be easily compared to a normative sample via the computer scoring system. This allows for interpretation of the quality of life score and comparison to a sample of healthy children. A normative sample is not available for comparison of pediatric patient-reported quality of life. Poor HRQOL has been defined as 2 SDs below the mean of the normative sample or a physical functioning or psychosocial health summary score <30 (2,3).

Respondent burden. Minimal burden; respondents generally answer 6 items per minute.

Administrative burden. Minimal burden; the administrator provides a brief introduction to the questionnaire, and then the authors indicate that completion takes ~1 minute for each of 6 items. Therefore, administration time varies from 5–25 minutes, depending on the number of items in the version being administered (i.e., 28, 50, or 87 items). No training is necessary for administration.

Translations/adaptations. The CHQ-PF50 and CHQ-PF28 have each been translated into 72 different languages, and the CHQ-CF87 has been translated into 25 different languages. A complete list of translations is available online at http://www.healthact.com/translation-chq.php.

Psychometric Information

Method of development. The CHQ was developed for children to assess HRQOL in a similar structure and methodology as that used by the Short Form 36 Health Survey (SF-36) (5). The scale was developed with parents of children ages 5–18 years with and without chronic health conditions using traditional item scaling analysis (6).

Acceptability. An examination of schoolchildren conducted by Raat and colleagues (7) indicated that <2% of data were missing on the CHQ-PF50 and up to 4% of items had nonunique answers. In another examination, Raat et al (5) examined the utility of the CHQ-PF28; results indicated that up to 1.7% of data were missing and up to 0.8% had nonunique answers. The authors have also compared the acceptability of the pencil and paper version of the CHQ-CF87 with an internet version. The internet version was found to yield fewer missing answers than the paper and pencil CHQ-CF87 (8).

Reliability. Studies have indicated that internal consistency for the CHQ-PF50 is good, with Cronbach's α for Dutch schoolchildren ranging from 0.39–0.96 for an average of 0.72 for the subscales (7). Additionally, Cronbach's α has been computed for US schoolchildren (0.66–0.94), children with asthma (0.67–-0.91), and children with attention deficit hyperactivity disorder (0.56–0.92) (9). The CHQ-PF28 has been found to demonstrate adequate internal consistency for the 2 summary scales, but the individual subscales demonstrate low internal consistency (5). Internal consistency for the CHQ-CF87 has been found to be adequate for the pencil and paper version and internet version, with Cronbach's α ranging from 0.69–0.92 (8).

Examination of test–retest reliability on the CHQ-PF50 indicated that intraclass correlations were significant for all but 2 scales, and test–retest means were not significantly different (7). Test–retest reliability for the CHQ-PF28 psychosocial summary scale was found to be excellent, but the individual scales were found to have low test–retest reliability (5).

Validity. *Construct validity.* The CHQ-CF87 has demonstrated good construct validity, with scores being lower for children with no chronic health conditions and higher for those with an increasing number of chronic conditions (8). These results were unaffected by mode of questionnaire administration (i.e., paper and pencil versus the internet). Further, exploratory and confirmatory factor analyses of the CHQ-PF50 with a sample of children and adolescents with various chronic illnesses, including juvenile idiopathic arthritis (JIA), have been conducted. These analyses suggest that the CHQ-PF50 demonstrates good construct validity for physical and psychosocial health constructs; however, the factor structure was observed to be different for children with chronic illnesses than for medically healthy children (10). Additionally, in a sample of children with systemic lupus erythematosus (SLE), the CHQ-PF50 has demonstrated good construct validity (1).

Convergent validity. Convergent validity for the CHQ-PF50 was examined using the Health Utilities Index in a sample of schoolchildren. Convergent validity was found to be acceptable, with correlations ranging from 0.21–0.49 for parallel domains on the questionnaires (7). Further, the CHQ-P50 has demonstrated good convergent validity with the Pediatric Quality of Life Inventory in a sample of children with SLE (1). The CHQ-PF28 was compared with a visual analog scale (VAS) to determine convergent validity. Convergent validity was found to be acceptable (0.15–0.50), and the VAS was found to correlate best with the general health perceptions subscale (0.50) (5).

Discriminant validity. Discriminant validity for the CHQ-PF50 was found to be moderate to strong when comparing children without a chronic medical condition to those with ≥2 chronic conditions, and when comparing those who had not attended a physician's appointment in the last year and those who had attended at least 3 times in the last year (7). The CHQ-PF28 demonstrated adequate discriminant validity, differentiating those children with a chronic health condition from those without (5).

Ability to detect change. In a sample of children with SLE, change in CHQ-PF50 physical health summary scores was observed to be consistent with changes in disease activity (1). However, the CHQ-PF50 psychosocial summary score and the CHQ-PF50 total score were observed to be less responsive to changes in health. Responsiveness of the CHQ-PF50 was also examined in a sample of Italian children with JIA (10). Similar to the pattern observed in children with SLE, the CHQ demonstrated good responsiveness to change in disease activity, with the physical health summary score evidencing better responsiveness than the total score or the psychosocial summary score. The responsiveness of the physical health summary score has also been examined independently in children with JIA. The CHQ was found to be sensitive to clinical change with a large standardized response mean for those who improved (0.96), small for those whose health was unchanged (0.16), and moderate for those whose health worsened (−0.60) (4).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CHQ has demonstrated adequate to good psychometric properties in a number of chronic illness populations. It also has both child and parent-proxy report versions, which allow for comparison of parent and child perceptions of child HRQOL. The CHQ is also available in a wide range of languages for cross-cultural comparison. The CHQ is easy to administer, and there is minimal respondent burden.

Caveats and cautions. Although the CHQ-PF50 has demonstrated good psychometric properties, the authors recommend using and interpreting summary scales on the CHQ-PF28 rather than individual scales, the latter of which have been found to have poor psychometrics. The CHQ may be confusing for some respondents because the item response options and recall periods vary by item. Further, the CHQ may only be used with parents and children ages 5–18 years and has not been validated for use with children ages <5 years.

Clinical usability. The CHQ requires minimal training for administration and scoring. It provides information on many discrete aspects of child HRQOL as well as overall scores; therefore, it may provide more detailed information for clinicians than other measures of HRQOL. Further, the CHQ can be mapped onto the SF-36, allowing for longitudinal measurement of HRQOL as patients transition from pediatric to adult care. However, the CHQ has several features that may limit use in a clinical setting. First, the completion time for the CHQ may inhibit clinic flow. Additionally, the CHQ may be expensive for regular use in a clinic. As the CHQ requires computer scoring, it does not allow clinicians to quickly review a patient's response and determine their level of HRQOL.

Research usability. The CHQ provides information regarding discrete aspects of child HRQOL. It is also very easy to administer and score. The internet version, which has shown similar psychometric properties to the traditional pencil and paper version, may be beneficial for research use because data entry will not be required. Additionally, a large normative sample is available for comparison across illness groups and with healthy children. Unfortunately, the varying item response options and recall periods may be confusing to children and their parents, so researchers should be available for clarification of items and verify that items were completed appropriately.

DISABKIDS CHRONIC GENERIC MEASURE (DCGM)

Description

Purpose. To assess health-related quality of life (HRQOL) in children and adolescents (ages 8–16 years) diagnosed with different chronic health conditions. The DISABKIDS, which was developed by the European DISABKIDS Group in 2002, is a modular measure and consists of both a generic form and 7 illness-specific forms. The following review will focus on the generic measures of HRQOL (refer to DISABKIDS Condition-Specific Measures for HRQOL, as it pertains to 7 different chronic illnesses).

Content. The DCGM consists of 3 domains of HRQOL: mental, social, and physical. Within each domain are 2 dimensions: independence (e.g., autonomy or living without impairments caused by the chronic health condition) and emotion (e.g., worries, concerns, or anger problems), social inclusion (e.g., acceptance of others, positive social relationships) and social exclusion (e.g., stigmatized, feeling left out), and limitation (e.g., functional limitations, perceived health status) and treatment (e.g., emotional impact of taking medication, receiving injections, taking insulin, etc.), respectively.

Number of items. The DCGM has 2 versions (long and short). The long version consists of 37 items (DCGM-37): mental (independence: 6 items, emotion: 7 items), social (social inclusion: 6 items, social exclusion: 6 items), and physical (limitation: 6 items, treatment: 6 items). The short version consists of 12 items and was derived from the DCGM-37.

Response options/scale. The DCGM-37 consists of ordinal scale items ranging from 1 (never) to 5 (always).

Recall period for items. Respondents are asked to refer back to the last 4 weeks.

Endorsements. The DISABKIDS Group.

Examples of use. Bullinger M, Schmidt S, Petersen C, and the DISABKIDS Group. Assessing quality of life of children with chronic health conditions and disabilities: a European approach. Int J Rehabil Res 2002;25:197–206.

Chaplin JE, Hanas R, Lind A, Tollig H, Wramner N, Lindblad B. Assessment of childhood diabetes-related quality-of-life in West Sweden. Acta Paediatrica 2008;98: 361–6.

Petersen C, Schmidt S, Bullinger M, and the DISABKIDS Group. Coping with a chronic pediatric health condition

and health-related quality of life. Eur Psychol 2006;11: 50–6.

Sandberg M, Johannson E, Bjork J, Wettergren L. Health-related quality of life related to school attendance in children on treatment for cancer. J Pediatr Oncol Nurs 2008; 25:265–74.

Practical Application

How to obtain. Interested parties are to complete a collaboration form (found online at http://www.disabkids.de/cms/licensing) and return it to the DISABKIDS Group. Following registration, the interested party will receive practical information (e.g., cost, versions) and login information to access questionnaires.

Method of administration. Two versions of the DCGM-37 are available: a child/adolescent self-report and a parent-proxy report. Both are paper and pencil questionnaires. A computer-assisted version is available.

Scoring. Hand scoring. Within each of the 6 subscales, item raw scores are summed and transformed into a scaled score ranging from 0 (poor HRQOL) to 100 (excellent HRQOL). Reference scores necessary for transformations are found in the DCGM manual. The subscales can also be combined to produce a general score of HRQOL (DCGM-37 total score). Missing values are to be substituted, if all but 1 item of each subscale is completed, by person-specific means based on his/her existing answers.

Score interpretation. The possible range for the DCGM total score is 37–185. Higher summed scores indicate better HRQOL.

Respondent burden. Minimal time to complete.

Administrative burden. Minimal training is necessary.

Translations/adaptations. Validated for use in the following languages: Dutch, English, French, German, Greek, and Swedish (12). Validation studies in Brazil and Mexico are currently being conducted.

Psychometric Information

Method of development. Development included focus groups of children and adolescents across Europe, in addition to parents and medical professionals, to identify aspects of HRQOL themes. Groups were classified by age, type of disease, and severity of disease. Results were used to derive items for the generic as well as the disease-specific modules (not discussed in this review).Three centers examined 3,027 statements for redundancy through a card-sort procedure. A total of 119 chronic generic items were selected to form the questionnaire for pilot testing.

Pilot studies were conducted to examine acceptability of the DCGM and its initial psychometric properties. The item-selection process following the pilot study resulted in a 56-item chronic generic questionnaire. Field tests were done to analyze the DCGM in a sample of 1,606 children and adolescents with a chronic condition. Participants were recruited from pediatric hospitals. There was equal representation across age ranges (4–7, 8–12, and 13–16 years), and the sample was primarily in the mild to moderate range of disease severity (38.7–50.7% of valid cases), although it also included more severe cases (10.6% of valid cases). Results from the field study provide the reference data reported in the manual.

Acceptability. No information.

Reliability. Internal consistency on the subscales (Cronbach's $\alpha = 0.70-0.87$) and test–retest reliability (intraclass correlation coefficient 0.71–0.83) is satisfactory across various chronic health conditions (12). In a sample of 117 Swedish children with cancer, internal consistency for the 6 subscales ranged from 0.71–0.87 (13).

Validity. *Content validity.* Items on the DCGM-37 were generated by focus groups, including children and adolescents with a chronic health condition, parents, and professionals (e.g., psychologists, physicians, and statisticians) (14). Analyses during pilot and field testing verified the grouping of items according to theoretical dimensions. Items with >5% of missing data, ceiling or floor effects of >60%, or absolute value of skewness of >2.0 were removed. When correlation coefficients between items were >0.8, redundant items were also discarded (12). A panel of experts classified all items according to age, type, and severity of disease.

Construct validity. In a sample of 1,153 children and adolescents (ages 8–16 years) with a chronic health condition (i.e., asthma, arthritis, epilepsy, cerebral palsy, diabetes mellitus, atopic dermatitis, cystic fibrosis), confirmatory factor analysis (root mean square of error approximation = 0.04, non-normed fit index = 0.95, comparative fit index = 0.95) supported a 6-factor structure for the final 37 items (12). Construct validity was further supported by satisfactory internal consistency on each of the 6 subscales.

Convergent validity. Simeoni and colleagues (12) found that the DCGM-37 was moderately associated with other already validated measures of HRQOL: Children's General Health Perceptions-Child Report (0.24–0.41), Functional Status-II (general health: 0.22–0.36), and Pediatric Quality of Life Inventory (physical: 0.23–0.70, emotional: 0.39–0.67, social: 0.19–0.64, and school: 0.35–0.59).

Discriminant validity. In the article by Simeoni et al (12), this was confirmed with girls and older adolescents reporting lower HRQOL compared to boys and younger children. Children from families with lower socioeconomic status and those with more severe diseases reported significantly lower HRQOL.

Ability to detect change. No information.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The DCGM is a reliable and valid measure of HRQOL and 6 specific dimensions across a variety of chronic health conditions. As one of the chronic health conditions included in development of the measure as well as pilot and field testing, it is well suited to assess HRQOL in children and adolescents with juvenile idiopathic arthritis (JIA). It has been validated in 6 languages thus far and has been utilized in different national and cultural contexts. The DCGM is easy to administer and score, with little training necessary.

Caveats and cautions. Research on the DCGM has largely focused on children and adolescents in European countries. As such, more studies are needed in the US to determine whether previous findings are generalizable to other children with chronic health conditions (e.g., JIA). The DCGM has examined HRQOL in JIA, but little is known about its applicability in other juvenile rheumatic diseases (e.g., lupus, dermatomyositis, spondylarthropathy). Furthermore, the DCGM does not provide information on how to interpret scores; therefore, it is difficult to know whether significant changes occur in child responses. The DCGM is also not useful for younger children (i.e., ages ≤8 years) with JIA.

Clinical usability. It is quick and easy to administer and score, limiting the burden to both respondents and clinicians. The psychometric properties and reference points of the DCGM indicate that it is a sound measure of HRQOL in JIA.

Research usability. Similarly, the DCGM is supported for its use in research with JIA, given that its development included this population. The measure is self-explanatory, allowing research participants to complete it with ease and without much assistance from researchers.

KINDL-R

Description

Purpose. To measure health-related quality of life (HRQOL) in healthy and ill children and adolescents (ages 4–16 years).

Content. The KINDL-R (15) consists of 24 items associated with 6 dimensions: physical well-being (e.g., illness, pain, fatigue), emotional well-being (e.g., boredom, loneliness, scared), self-esteem (e.g., pride, feeling on top of the world), family (e.g., relationship with parents, conflict at home), friends (e.g., getting along, feeling different from others), and everyday functioning in school (e.g., enjoying class, worrying about the future). Disease is an optional subscale (e.g., illness uncertainty, parent overprotection, missing school) that can be added in the case of prolonged illness or hospitalization. Disease-specific modules are available for children with obesity, bronchial asthma, atopic dermatitis, and diabetes mellitus.

Number of items. The KINDL-R consists of 24 items, with each subscale containing 4 items.

Response options/scale. Responses are on a 5-point ordinal scale from 1 (never) to 5 (all of the time).

Recall period for items. Respondents are asked to refer to the past week.

Endorsements. No information.

Examples of use. Ertan P, Yilmaz O, Caglayan M, Sogut A, Aslan S, Yuksel H. Relationship of sleep quality and quality of life in children with monosymptomatic enuresis. Child Care Health Dev 2008;35:469–74.

Milde-Busch A, Heinrich S, Thomas S, Kuhnlein A, Radon K, Straube A, et al. Quality of life in adolescents with headache: results from a population-based survey. Cephalagia 2010;30:713–21.

Muller-Godeffroy E, Lehmann H, Kuster RM, Thyen U. Quality of life and psychosocial adaptation in children and adolescents with juvenile idiopathic arthritis and reactive arthritis. J Rheumatol 2005;64:177–87.

Ravens-Sieberer U, Bullinger M. Assessing the health related quality of life in chronically ill children with the German KINDL: first psychometric and content-analytical results. Qual Life Res 1998;7:399–407 (15).

Ravens-Sieberer U, Erhart M, Wille N, Bullinger M. Health-related quality of life in children and adolescents in Germany: results of the BELLA study. Eur Child Adolesc Psychiatry 2008;17 Suppl:148–56.

Wille N, Erhart M, Petersen C, Ravens-Sieberer U. The impact of overweight and obesity on health-related quality of life in childhood: results from an intervention study. BMC Public Health 2008;8:421–9.

Practical Application

How to obtain. The KINDL-R may be used with permission from the developers (www.kindl.org). The manual, computer software, and questionnaires are free for all nonprofit or research institutions only, under the condition that a user form is completed. No other cost information is provided.

Method of administration. Three versions of the KINDL-R are available as self-report measures for different age groups: Kiddy-KINDL-R (ages 4–7 years; interview format), Kid-KINDL-R (ages 8–12 years), and Kiddo-KINDL-R (ages 13–16 years). It is also available in 2 parent-proxy versions (ages 4–7 years and 8–16 years). A shorter, 12-item version of the KINDL-R and a computer-assisted, touch screen version (CAT-Screen) are also available.

Scoring. The KINDL-R is scored with computer scoring software. Briefly, 10 items are reversed before being summed to reach 6 subscale scores (physical well-being, emotional well-being, self-esteem, family, friends, and school). If necessary, an additional subscale score for disease can be added. Subscales can be combined for a total score, or they can be transformed to values between 0 and 100. Scoring the parent version follows the same general steps. Instructions for common coding problems include: if 2 responses are marked for a single question and these responses are adjacent to one another, then 1 response is chosen according to a random procedure and entered; if 2 responses are marked for a single question and these responses are not adjacent to one another, then the item is coded as a missing value; and if 3 or more responses are marked for a single question, the item is coded as a missing value. The algorithm on the computer software replaces any missing values by an estimate made specifically for that person, provided that the respondent answered at least 70% of the items on the subscale.

Score interpretation. Higher scores on the KINDL-R indicate better HRQOL.

Respondent burden. Minimal; <15 minutes to complete, and the KINDL-R is self-explanatory.

Administrative burden. Minimal training is necessary for administration. Scoring requires training in SPSS software.

Translations/adaptations. The original KINDL was developed in German and is also available in English, Dutch, French, Greek, Italian, Norwegian, Russian, Spanish, Swedish, and Turkish. The Turkish, English, and Spanish KINDL have been validated (16–18).

Psychometric Information

Method of development. No information.

Acceptability. Floor and ceiling effects are <10%.

Reliability. In a sample of 1,050 children and adolescents (mean age 12.6 years) with bronchial asthma, atopic dermatitis, or obesity who were recruited from 7 German rehabilitation clinics, internal consistency for the KINDL-R subscales was satisfactory (0.63–0.76) and good for the total score (0.84) (19). For the parent version, internal consistency was satisfactory for the subscales (0.62–0.81) and excellent for the total score (0.89) (16).

In a study by Erhart and colleagues (20), HRQOL using the KINDL-R and KINDL proxy version was examined among 6,813 German children and adolescents (ages 11–17 years; 17.5% classified as having a chronic health condition) and their parents. Internal consistency for the total score was slightly higher for the parent report than the child report (0.86 versus 0.82) (20).

Validity. *Content validity.* No information.

Construct validity. Using confirmatory factor analysis, the KINDL-R had an acceptable fit to the 6-dimensional model for the parent (root mean square of error approximation [RMSEA] = 0.07, comparative fit index [CFI] = 0.95) and the child (RMSEA = 0.06, CFI = 0.93) (20). Construct validity was further supported by satisfactory internal consistency on each of the 6 dimensions (child range 0.53–0.72, parent range 0.62–0.72).

Convergent validity. Ravens-Sieberer and colleagues (19) found the KINDL-R to be associated with other measures of HRQOL, including the Children's Health Questionnaire (general well-being: r = 0.7), Short Form 36 Health Survey (SF-36; vitality: 0.64, emotional well-being: 0.64), and Life Satisfaction Questionnaire adapted for children (life satisfaction: 0.69). Erhart and colleagues (20) reported similar findings with associations between the child and parent KINDL-R and the Strength and Difficulties Questionnaire (SDQ; child range 0.33–0.49, parent range 0.44–0.53).

Discriminant validity. In the same study by Erhart and colleagues (20), discriminant validity was indicated by low correlations between the KINDL-R and opposing dimensions of the SDQ. Regarding the ability to discriminate between healthy children and those with a chronic health condition, the child-report form exhibited small effect sizes (0.04–0.27) and the parent-report form had medium effect sizes (0.20–0.56) (20). Parent- and child-report total scores and physical well-being scores had small effect sizes (parent: 0.31 and 0.26, respectively, and child: 0.25 and 0.18, respectively), whereas the child report yielded large effect sizes for the impact of obesity on the dimensions of self-esteem (0.19), friends (0.28), and school-related well-being (0.23) (20).

Ability to detect change. No information.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The KINDL-R is a flexible, modular, and psychometrically sound measure of HRQOL for children and adolescents with and without a chronic health condition. A few primary advantages of this measure are its self- and parent-report measures, which may be used to assess concordance rates, as well as age-specific versions to account for the changes that take place over the course of the child's development.

Caveats and cautions. Although this measure has been studied in various chronic health conditions (e.g., diabetes mellitus and cerebral palsy), less information is available on juvenile rheumatic diseases (JRDs). This limitation warrants future studies examining the generalizability to other chronic illnesses. Furthermore, the KINDL-R does not allow use throughout the entire pediatric age range, and is not appropriate for children ages <4 years with a JRD. Although the psychometric properties have been examined in some of the translated versions of the KINDL, the reliability and validity for several versions have not been investigated. Lastly, how to interpret scores on the KINDL-R is unknown, making it difficult to assess changes in HRQOL in children with JRDs.

Clinical usability. The KINDL-R requires little time and effort on the part of the respondent, whether a child or parent. It has wide applicability in various settings such as community or clinical mental health and medical settings. Scoring, however, requires that the clinician have some knowledge of SPSS software. Therefore, the KINDL-R may not be the quickest measure of HRQOL in a clinical setting. Additionally, little research has been done on the appropriateness of using the KINDL-R in pediatric rheumatology clinics.

Research usability. This measure can be completed by several research participants in a short amount of time and offers a great deal of information regarding overall HRQOL, in addition to 6 specific domains. Again, scoring may be time consuming.

PEDIATRIC QUALIFY OF LIFE INVENTORY (PEDSQL) 4.0 GENERIC CORE SCALES

Description

Purpose. To measure health-related quality of life (HRQOL) in children and adolescents ages 2–18 years. This measure consists of child report (ages 5–18 years) and parent report (ages 2–18 years) of the child's HRQOL, and can be used with healthy children and those with acute and chronic health conditions. PedsQL 4.0 is the fourth and current version. The PedsQL 4.0 Generic Core Scales were specifically designed to measure the core health dimensions outlined by the World Health Organization.

Content. Physical, emotional, social, and school functioning. Specifically, questions inquire about problems related to child health, activities, feelings, getting along with others, and school.

Number of items. 23 items for the total scale score: 8 items for physical health summary score and 15 items for psychosocial health summary.

Response options/scale. For children ages 8–18 years and parent-proxy report formats, items are rated on a 5-point ordinal scale to indicate how much the child has problems with various areas of functioning, ranging from 0 (never) to 4 (almost always). For younger children, the ordinal scale is reworded and simplified to a 3-point scale: 0 (not at all a problem), 2 (sometimes a problem), and 4 (a lot of a problem).

Four subscales, including physical functioning (8 items), emotional functioning (5 items), social functioning (5 items), and school functioning (5 items), contribute to 3 summary scores: total scale score (all subscales), physical health summary score (physical functioning scale only), and psychosocial health summary (emotional, social, and school functioning scales combined).

Recall period for items. 1 month.

Endorsements. No information.

Examples of use. Brunner HI, Taylor J, Britto MT, Corcoran MS, Kramer SL, Melson PG, et al. Differences in disease outcomes between Medicaid and privately insured children: possible health disparities in juvenile rheumatoid arthritis. Arthritis Rheum 2006;55:378–84.

Brunner HI, Higgins GC, Wiers K, Lapidus SK, Olson JC, Onel K, et al. Health-related quality of life and its relationship to patient disease course in childhood-onset systemic lupus erythematosus. J Rheumatol 2009;36:1536–45 (1).

Moorthy LN, Harrison MJ, Peterson M, Onel KB, Lehman TJ. Relationship of quality of life and physical function measures with disease activity in children with systematic lupus erythematosus. Lupus 2005;14:280–7.

Ringold S, Wallace CA, Rivara FP. Health-related quality of life, physical function, fatigue, and disease activity in children with established polyarticular juvenile idiopathic arthritis. J Rheumatol 2009;36:1330–6.

Robinson RF, Nahata MC, Hayes JR, Rennebohm R, Higgins G. Quality of life measurements in juvenile rheumatoid arthritis patients treated with Etanercept. Clin Drug Investig 2003;23:511–8.

Sandstrom MJ, Schanberg LE. Peer rejection, social behavior, and psychological adjustment in children with juvenile rheumatic disease. J Pediatr Psychol 2004;29: 29–34.

Sawyer MG, Whitham JN, Roberton DM, Taplin JE, Varni JW, Baghurst PA. The relationship between health-related quality of life, pain, and coping strategies in juvenile idiopathic arthritis. J Rheumatol 2004;43:325–30.

Trapanotto M, Giorgino D, Zulian F, Benini F, Varni JW. The Italian version of the PedsQL in children with rheumatic diseases. Clin Exp Rheumatol 2009;27:373–80.

Practical Application

How to obtain. The PedsQL scales, modules, and translations are protected by copyright. Upon accepting the user agreement, a single copy of the measure can be obtained online at http://pedsql.org/pedsql12.html. Individuals, organizations, or institutions wishing to order the PedsQL should contact: Christelle Berne, Mapi Research Institute, 27, rue de la Villette, 69003 Lyon, France; e-mail: cberne@mapi.fr; telephone: +33 (0) 472 13 66 67.

The PedsQL is free to individuals conducting nonfunded academic research; however, the cost for funded academic research and large noncommercial organization research and evaluations (e.g., states, nations, hospitals, health care systems) and commercial studies can vary widely ($1,000–$20,000). Fee calculators can be found online at http://www.pedsql.org/conditions.html.

Method of administration. Parents, children (ages 8–12 years), and adolescents (ages 13–18 years) may self-administer the PedsQL after instructions from the administrator. For younger children (ages 5–7 years) or if the child or adolescent is unable to self-administer the PedsQL (e.g., due to illness, fatigue, reading difficulties), the measure should be read aloud. General protocol and administration guidelines (including a script) are available online at http://www.pedsql.org/pedsqladmin.html.

Scoring. Items are reverse scored and linearly transformed to a 0–100 scale, so that higher scores indicate better HRQOL. To reverse score, transform the scale items to 0–100 as follows: 0 = 100, 1 = 75, 2 = 50, 3 = 25, and 4 = 0. To create scale scores, the mean is computed by totaling the item scores and dividing by the number of items answered (this accounts for missing data). If >50% of the items in the scale are missing, it is recommended that the scale score not be computed. Imputing the mean of the completed items in a scale when 50% or more are completed is generally the most unbiased method. To create the psychosocial health summary score, the mean is computed as the sum of the items over the number of items answered in the emotional, social, and school functioning subscales. The physical health summary score is the same as the physical functioning subscale score. To create the total scale score, the mean is computed as the sum of all of the items over the number of items answered on all of the scales. Computer scoring is not necessary.

Score interpretation. The range on subscales and the overall scale is 0–100, with lower scores indicating poorer HRQOL and higher scores indicating better HRQOL. When examining the total scale, scores of 4.4 and 4.5 are considered to be minimal clinically meaningful differences on the child self-report and parent-proxy report, respectively (21).

Respondent burden. Minimal; questions are written at a third- to sixth-grade reading level, and the entire questionnaire takes <4 minutes to complete.

Administrative burden. Minimal; the administrator provides a brief introduction to the questionnaire and then administration of the PedsQL takes <4 minutes, even when reading of the questionnaire is required. Scoring also takes a minimal amount of time (several minutes) and effort. No training necessary for administration.

Translations/adaptations. The PedsQL Generic Core questionnaire has been linguistically validated for children and adolescents (ages 2–18 years) and parents in the following languages: Belgium Dutch, Belgium French, Portuguese for Brazil, French for Canada, Croatian, Czech, Danish, French, German, Hungarian, Hebrew, Italian, Latvian, Lithuanian, Spanish for Mexico, Norwegian, Urdu for Pakistan, Spanish for Peru, Polish, Portuguese, Russian, Slovakian, Slovenian, Spanish, and Swedish. It has also been translated, but not formally validated, into a variety of other languages. A complete list of translations is available online at http://pedsql.org/translations.html.

Psychometric Information

Method of development. The original PedsQL 1.0 was empirically derived from data collected from 291 pediatric patients with cancer and their parents at various stages of treatment. It was designed as a generic HRQOL instrument to be utilized across diverse pediatric populations. The PedsQL 2.0 and 3.0 included additional constructs and items, a more sensitive scaling range, and a broader age range for child self-report and parent-proxy report (22). The PedsQL 4.0 Generic Core Scales have resulted from this process and were specifically designed to measure the core health dimension outlined by the World Health Organization.

In the initial field trail, the PedsQL 4.0 Generic Core Scales were administered to 963 children and 1,629 parents (23). Later, the psychometric properties of the PedsQL 4.0 were tested in a group of children (n = 231) with rheumatoid arthritis (e.g., juvenile idiopathic arthritis [JIA; pauciarticular, polyarticular, and systemic subtypes], systemic lupus erythematosus, juvenile fibromyalgia, spondylarthritis, other rheumatic diseases) and their parents (22). Psychometric statistics provided below are from the investigation of children with juvenile rheumatic diseases.

Acceptability. Recent studies suggest that 0.7% of child report and 3% of parent report data were missing. Items about school were most frequently skipped, suggesting that these were not completed when children did not attend school during the previous month (when given in the summer) (22).

Reliability. Internal consistencies for the total scale score were as follows: child self-report Cronbach's α = 0.91, parent-proxy report Cronbach's α = 0.93; physical health summary scale score: child self-report Cronbach's α = 0.87, parent-proxy report Cronbach's α = 0.89; and psychosocial health summary scale score: child self-report Cronbach's α = 0.86, parent-proxy report Cronbach's α = 0.90.

Validity. Construct validity was determined by comparing scale scores across children with juvenile rheumatic diseases (JRDs) and healthy children, because these groups are known to differ in HRQOL. For every comparison (i.e., self- and parent-report of total score, physical health summary score, psychosocial health summary score, emotional functioning, social functioning, and school functioning), a statistically significant difference existed when comparing healthy children to children with JRDs (22). In other words, healthy children had higher PedsQL 4.0 scores (suggesting better HRQOL) than children with rheumatic diseases.

Ability to detect change. The responsiveness of the PedsQL was demonstrated through a longitudinal analysis of change within participants with JRDs for whom change was expected as a result of an intervention (22). For both child self-report and parent-proxy report, the PedsQL 4.0 Generic Core Total and summary scale scores increased progressively from visit 1 through visit 3. Effect sizes for the difference between visit 1 and 2 for child self-report (d = 0.34) and parent proxy (d = 0.27) were in the small range, while the effect sizes for the difference between visit 1 and 3 for child self-report (d = 0.92) and parent-proxy report (d = 0.71) were in the medium to large effect size range.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PedsQL Generic Core Scales are widely used in a variety of pediatric patient populations (i.e., JIA, systemic lupus erythematosus) and with healthy children. The measure is brief, developmentally appropriate for a broad range of ages, reliable, valid, responsive, and translated into many languages. Additionally, disease-specific modules (not discussed in this review) can also be added to the Generic Core Scales to provide both disease-specific and general measures of quality of life.

Caveats and cautions. Because the PedsQL asks children and parents to remember information from the past month, children and parents may have difficulty completing school-related questions if the child has not recently been in school.

Clinical usability. The PedsQL 4.0 Generic Core Scales are a practical measure for clinicians. In a short amount of time (~4 minutes), physicians and clinicians can gather general information about the child's physical, emotional, social, and school functioning.

Research usability. Overall, the PedsQL 4.0 Generic Core Scales can be used as an excellent measure of general HRQOL. Additionally, the measure can be self-administered and understood by most adults and children.

QUALITY OF MY LIFE QUESTIONNAIRE (QOML)

Description

Purpose. To assess quality of life (QOL) and health-related quality of life (HRQOL) as 2 separate constructs in children and adolescents.

Content. Two visual analog scales (VAS) and a categorical measure of change in QOL.

Number of items. 3 items: the QOL and HRQOL VAS and a categorical item assessing change in QOL since the previous visit.

Response options/scale. Children and parent-proxy reporters each complete 2 VAS. The QOL VAS asks "Overall, my life is . . . ," and the HRQOL VAS asks "Considering my health, my life is. . . ." Respondents are asked to record their responses on a 100-mm VAS for each question stem, which ranges from 0 (the worst) to 100 (the best). Respondents also provide a categorical response to the question, "Since the last time I was here, my life is. . . ." The item is rated on a 5-point ordinal scale ranging from much worse to much better.

Recall period for items. Current.

Endorsements. No information.

Examples of use. Dempster H, Porepa M, Young N, Feldman BM. The clinical meaning of functional outcome scores in children with juvenile arthritis. Arthritis Rheum 2001;44:1768–74.

Gong GW, Young NL, Dempster H, Porepa M, Feldman BM. The Quality of My Life questionnaire: the minimal clinically important difference for pediatric rheumatology patients. J Rheumatol 2007;34:581–7 (24).

Oen K, Tucker L, Huber AM, Miettunen P, Scuccimarri R, Campillo S, et al. Predictors of early inactive disease in a juvenile idiopathic arthritis cohort: results of a Canadian multicenter, prospective inception cohort study. Arthritis Rheum 2009;61:1077–86.

Singh-Grewal D, Schneiderman-Walker J, Wright V, Bar-Or O, Beyene J, Selvadurai H, et al. The effects of vigorous exercise training on physical function in children with arthritis: a randomized, controlled, single-blinded trial. Arthritis Rheum 2007;57:1202–10.

Stephens S, Feldman BM, Bradley N, Schneiderman J, Wright V, Singh-Grewal D, et al. Feasibility and effectiveness of an aerobic exercise program in children with fibromyalgia: results of a randomized controlled pilot trial. Arthritis Rheum 2008;59:1399–406.

Practical Application

How to obtain. The QoML is available in the appendices of the articles by Gong et al (24) and Feldman et al (25).

Method of administration. Parents and children (ages 8–12 years) may self-administer the QoML after instructions from the administrator.

Scoring. The VAS are scored by measuring the length (in mm) of the line between the left anchor and the respondent's mark on the line.

Score interpretation. The response range is from 0–100, with higher scores suggesting better QOL.

Respondent burden. Minimal burden; respondents complete 3 items, and the questionnaire takes <5 minutes to complete.

Administrative burden. Minimal burden; the administrator provides a brief introduction to the questionnaire, and then administration takes <5 minutes, even when reading of the questionnaire is required. Scoring also takes minimal time and effort. No training is necessary for administration.

Translations/adaptations. None.

Psychometric Information

Method of development. The QoML was developed in a pediatric rheumatology sample as a measure of both QOL and HRQOL. The scale is based upon other VAS that have been widely used to assess for QOL in adult cancer samples. The QoML was pretested on 10 children ages 4–12 years; 6 of the children were rheumatology clinic patients and 4 were healthy children. Face validity was assessed by 11 pediatric rheumatology professionals and found to be good to strong. The questionnaire was piloted on 122 pediatric rheumatology patients of various diagnoses, including pauciarticular juvenile arthritis, spondylarthropathy, fibromyalgia, and osteomyelitis, ranging in age from 10 months to 18 years. The patients and their parents independently determined at what age the child was able to complete the questionnaire for himself/herself or at what age a parent would need to provide proxy report. Results indicated that QOL and HRQOL were viewed by respondents as related yet discrete constructs.

Acceptability. No information.

Reliability. No information.

Validity. Convergent construct validity was determined by Feldman and colleagues (25) and Gong and colleagues (24) in pediatric rheumatology samples by comparing respondents' scores on the QOL and HRQOL VAS and the Childhood Health Assessment Questionnaire, a traditional measure of health status. For both examinations, results indicated that convergent construct validity was good, as the relationships between the scales and disease variables (i.e., disease severity, disability, morning stiffness, pain) were as expected. Criterion validity was not assessed.

Ability to detect change. With regard to the responsiveness of the QoML, or ability to detect change over time, Gong and colleagues (24) conducted an examination to determine the minimum clinically important difference (MCID) in QOL and HRQOL for pediatric rheumatology patients and their parents. MCID was determined by asking patients and their parents to provide VAS responses for their current QOL and HRQOL and 2 hypothetical situations. One situation suggests that the child's health had improved "just enough to make a difference," and the other suggests that the child's health had gotten worse "just enough to make a difference" in his/her QOL and HRQOL. The authors were able to provide numerical values of responsiveness for the QoML in the pediatric rheumatology population. MCID for improvement was 7 mm and 11 mm for QOL and HRQOL, respectively, and MCID for deterioration was −33 mm and −38 mm for QOL and HRQOL, respectively (24).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The QoML was developed in a pediatric rheumatology sample, making it appropriate for other pediatric rheumatology populations. The questionnaire is quick and easy to administer and requires minimal reading on the part of the child or parent. Unlike most measures of QOL, the QoML provides an assessment of both the child's QOL and HRQOL.

Caveats and cautions. The QoML has limited psychometric data and is still in the process of being validated. Currently, age limitations do not allow the QoML to be used throughout the entire pediatric age range, with parents able to report on children as young as 10 months and children able to report on their own QOL as young as 4 years. Additionally, unlike measures of QOL with more items and subscales, the QoML does not provide domain-specific information about what factors (e.g., physical, emotional, social, behavioral, academic) are contributing to a child's change in QOL or HRQOL.

Clinical usability. The QoML is an easy measure to administer and score. Clinicians can quickly determine a child's QOL or HRQOL by viewing their completed VAS and their report of qualitative change since the subsequent visit. It may be very useful for a clinician who is interested in a quick estimate of a child's QOL and change since the previous visit.

Research usability. The QoML can be completed quickly and easily by both child and parent research participants. The measure is available for free in the appendices of the authors' studies. Unlike most measures of pediatric QOL, the QoML provides a report of both QOL and HRQOL.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Brunner HI, Higgins GC, Wiers K, Lapidus SK, Olson JC, Onel K, et al. Health-related quality of life and its relationship to patient disease course in childhood-onset systemic lupus erythematosus. J Rheumatol 2009;36:1536–45.
2. Gutierrez-Suarez R, Pistorio A, Cespedes Cruz A, Norambuena X, Flato B, Rumba I, et al. Health-related quality of life of patients with juvenile idiopathic arthritis coming from 3 different geographic areas: the PRINTO multinational quality of life cohort study. Rheumatology (Oxford) 2007;46:314–20.
3. Oliveira S, Ravelli A, Pistorio A, Castell E, Malattia C, Prieur AM, et al. for the Pediatric Rheumatology International Trials Organization. Proxy-reported health-related quality of life of patients with juvenile idiopathic arthritis: the Pediatric Rheumatology International Trials Organization multinational quality of life cohort study. Arthritis Rheum 2007;57:35–43.
4. Selvaag AM, Flato B, Lien G, Sorskaar D, Vinje O, Forre O. Measuring health status in early juvenile idiopathic arthritis: determinants and responsiveness of the Child Health Questionnaire. J Rheumatol 2003; 30:1602–10.
5. Raat H, Botterweck AM, Landgraf JM, Hoogeveen WC, Essink-Bot M. Reliability and validity of the short form of the Child Health Questionnaire for parents (CHQ-PF28) in large random school based and general population samples. J Epidemiol Community Health 2005;59:75–82.
6. Asmussen L, Olson LM, Grant EN, Landgraf JM, Fagan J, Weiss KB. Use of the Child Health Questionnaire in a sample of moderate and low-income inner-city children with asthma. Am J Respir Crit Care Med 2000;162:1215–21.
7. Raat H, Bonsel GJ, Essink-Bot M, Landgraf JM, Gemke RJ. Reliability and validity of comprehensive health status measures in children: the Child Health Questionnaire in relation to the Health Utilities Index. J Clin Epidemiol 2002;55:67–76.
8. Raat H, Mangunkusumo RT, Landgraf JM, Kloek G, Brug J. Feasibility, reliability, and validity of adolescent health status measurement by the Child Health Questionnaire Child Form (CHQ-CF): internet administration compared with standard administration. Qual Life Res 2007;16: 675–85.
9. Landgraf JL, Abetz L, Ware JE. The CHQ user's manual. Boston: The Health Institute, New England Medical Center; 1996.
10. Drotar D, Schwartz L, Palermo TM, Burant C. Factor structure of the Child Health Questionnaire-Parent Form in pediatric populations. J Ped Psychol 2006;31:127–38.
11. Moretti C, Viola S, Pistorio A, Magni-Manzoni S, Ruperto N, Martini A, et al. Relative responsiveness of condition specific and generic health status measures in juvenile idiopathic arthritis. Ann Rheum Dis 2005; 64:257–61.
12. Simeoni MC, Schmidt S, Muehlan H, Debensason D, Bullinger M, and the DISABKIDS Group. Field testing of a European quality of life instrument for children and adolescents with chronic conditions: the 37-item DISABKIDS Chronic Generic Module. Qual Life Res 2007;16: 881–93.
13. Sandeberg MA, Johansson EM, Hagell P, Wettergren L. Psychometric properties of the DISABKIDS Chronic Generic Module (DCGM-37) when used in children undergoing treatment for cancer. Health Qual Life Outcomes 2010;8:109–26.
14. Petersen C, Schmidt S, Power M, Bullinger M, and the DISABKIDS Group. Development and pilot-testing of a health-related quality of life chronic generic module for children and adolescents with chronic health conditions: a European perspective. Qual Life Res 2005;14: 1065–77.
15. Ravens-Sieberer U, Bullinger M. Assessing the health related quality of life in chronically ill children with the German KINDL: first psychometric and content-analytical results. Qual Life Res 1998;7:399–407.
16. Eser E, Yuksel H, Baydur H, Erhart M, Saatli G, Cengiz-Ozyurt B, et al. The psychometric properties of the new Turkish generic health-related quality of life questionnaire for children (Kid-KINDL). Turkish J Psychiatry 2008;19:409–17.
17. Fernandez-Lopez J, Fernandez Fidalgo M, Cieza A, Ravens-Sieberer U. Measuring health-related quality of life in children and adolescents: preliminary validation and reliability of the Spanish version of the KINDL questionnaire. Aten Primaria 2004;33:434–42.
18. Wee H, Lee W, Ravens-Sieberer U, Erhart M, Li S. Validation of the English version of the KINDL generic children's health-related quality of life instrument for an Asian population: results from a pilot test. Qual Life Res 2005;14:1193–200.
19. Ravens-Sieberer U, Redegeld M, Bullinger M. Lebensqualitat chronisch kranker KINDER im verlauf der statioaren rehabilitation. In: Neuser J, de Bruin JT, editors. Verbindung und veränderung im fokus der medizinischen psychologie. Lengerich (Germany): Pabst Science Publishers; 2000. p. 89–91.
20. Erhart M, Ellert U, Kurth B, Ravens-Sieberer U. Measuring adolescents' HRQOL via self reports and parent proxy reports: an evaluation of the psychometric properties of both versions of the KINDL-R instrument. Health Qual Life Outcomes 2009;7:77–88.
21. Varni JW, Burwinkle TM, Seid M, Skarr D. PedsQL 4.0 as a pediatric population health measure: feasibility, reliability, and validity. Ambul Pediatr 2003;3:329–41.
22. Varni JW, Seid M, Knight TS, Burwinkle TM, Brown J, Szer IS. The PedsQL in pediatric rheumatology: reliability, validity, and responsiveness of the Pediatric Quality of Life Inventory Generic Core Scales and Rheumatology Module. Arthritis Rheum 2002;46:714–25.
23. Varni JW, Seid M, Kurtin PS. PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 generic core scales in healthy and patient populations. Med Care 2001;39:800–12.
24. Gong GW, Young NL, Dempster H, Porepa M, Feldman BM. The Quality of My Life questionnaire: the minimal clinically important difference for pediatric rheumatology patients. J Rheumatol 2007;34:581–7.
25. Feldman BM, Grundland B, McCullough L, Wright V. Distinction of quality of life, health-related quality of life, and health status in children referred for rheumatologic care. J Rheumatol 2003;27:226–33.

Summary Table for General Pediatric Measures of Quality of Life*

Scale	Purpose/ content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
CHQ	To measure HRQOL in children (ages 5–18)	Parents and children (ages 10–18) may self-administer after a brief introduction	Minimal; 5–25 minutes (depending on version administered)	Minimal; no training necessary	0 (worst possible health status) to 100 (best possible health status) Normative sample available for comparison	CHQ-PF50: α = 0.66–0.92; test–retest means not significantly different CHQ-PF28: test–retest reliability was excellent CHQ-CF87: α = 0.69–0.92	CHQ-PF50: construct good, convergent acceptable with HUI and good with PedsQL, discriminant moderate to strong CHQ-PF28: discriminant adequate, convergent acceptable with VAS CHQ-CF87: construct good	CHQ-PF50: good responsiveness in samples of children with JIA and SLE; physical health summary score is most responsive to change in disease activity	Good psychometric properties in many chronic illness populations Child and parent-proxy versions Available in a wide variety of languages Normative sample available for comparison	CHQ-PF28: individual scales have poor psychometrics May be confusing because the item response options and recall periods vary by item Only validated in children ages 5–18 years
DCGM	To assess HRQOL in children (ages 8–16 years)	Child self-report and parent-proxy report Paper and pencil and computer-assisted versions	Minimal	Minimal training is necessary	37–185, with higher summed scores indicating better HRQOL	α = 0.70–0.87; test–retest is satisfactory	Construct: supported by satisfactory internal consistency on each of 6 subscales	No information	Reliable and valid Developed on populations of children with JIA Validated in 6 languages	Little research has been conducted on children and adolescents in the US Little research on its use with other JRDs
KINDL-R	To measure HRQOL in children (ages 4–16 years)	Self-report interview (ages 4–7 years) Self-report (ages 8–12 and 8–16 years) Parent proxy Paper and pencil and computer-assisted versions available	Minimal; <15 minutes to complete Self-explanatory	Minimal training is necessary Scoring requires training in SPSS software	0–100, with higher scores indicating better HRQOL	α = 0.84–0.89 for overall score α = 0.63–0.76 for subscales child report α = 0.62–0.81 for parent subscales	Construct: acceptable fit for the 6-dimensional model for parent and child	No information	Flexible, modular, psychometrically sound measure Concordance rates can be calculated Several age-specific versions are available	Little information available on using this measure in populations of children with JRDs
PedsQL 4.0 Generic Core Scales	To measure HRQOL in children (ages 2–18 years)	Child self-report and parent-proxy report are self-administered Paper and pencil questionnaire	Minimal; <4 minutes 3rd–6th-grade reading level	Minimal No training necessary Administrator provides a brief introduction Scoring takes minimal time and effort	0–100, with higher scores indicating better HRQOL; ~4.5 indicates minimum clinically important difference	Total scale score child self-report (α = 0.91), parent proxy (α = 0.93); scales (α = 0.86–0.90)	Construct: children with JRDs had lower scores (worse HRQOL) than healthy children	Longitudinal analysis demonstrated an increase in scores following an intervention	Brief, valid, reliable, and developmentally appropriate measure Translated into many languages Disease-specific modules are available	Children and parents may have difficulty completing school-related questions if the child has not been in school in the past month
QoML	To assess QOL and HRQOL as 2 separate constructs in children (ages 10 months to 18 years)	Child self-report and parent-proxy report are self-administered Paper and pencil questionnaire	Minimal; only 3 items, takes <5 minutes to complete	Minimal No training necessary Administrator provides a brief introduction Scoring takes minimal time and effort	VAS are scored by measuring the length between anchor and mark 0–100, with higher scores indicating better QOL	No information	Construct validity is good; relationship between scales and disease variables were as expected	Responses to hypothetical situations using "just enough change to make a difference" suggest that change can be observed on the VAS	Developed with a pediatric rheumatology sample Quick and easy questionnaire Requires minimal reading Assesses both QOL and HRQOL	Limited psychometric data available Does not provide information about factors contributing to a child's QOL or HRQOL

* CHQ = Child Health Questionnaire; HRQOL = health-related quality of life; CHQ-PF50 = CHQ long parent-report questionnaire; CHQ-PF28 = CHQ short parent-report questionnaire; CHQ-CF87 = CHQ child-report questionnaire; HUI = Health Utilities Index; PedsQL = Pediatric Quality of Life Inventory; VAS = visual analog scale; JIA = juvenile idiopathic arthritis; SLE = systemic lupus erythematosus; DCGM = DISABKIDS Chronic Generic Measure; JRDs = juvenile rheumatic diseases; QoML = Quality of My Life Questionnaire.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S467–S472
DOI 10.1002/acr.20561
© 2011, American College of Rheumatology

Measures of Anxiety

State-Trait Anxiety Inventory (STAI), Beck Anxiety Inventory (BAI), and Hospital Anxiety and Depression Scale-Anxiety (HADS-A)

LAURA J. JULIAN

INTRODUCTION

This review covers commonly used measures of anxiety. For this review, the author included measures that were 1) measures of general measures of anxiety and severity of anxiety symptoms, 2) administered by self-report, 3) used in rheumatologic populations, and 4) has evidence of adequate psychometric data. To maintain brevity, the majority of the measures reviewed here were selected to provide broad coverage of general symptoms of anxiety, and measures were excluded if they are intended to identify or characterize a specific Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) anxiety disorder (1). Specifically, this author excluded measures typically used to evaluate diagnostic criteria or features of specific anxiety disorders, such as panic disorder, obsessive-compulsive disorder, posttraumatic stress disorder, and others. In addition, broader measures of psychiatric distress, including the Symptom Checklist-90, the General Health Questionnaire, and the Medical Outcomes Study Short Form 36 are not included in this review since they are included elsewhere in this special issue.

However, subscales that have been used frequently in rheumatology as "stand-alone" measures, such as the anxiety scale of the Hospital Anxiety and Depression Scale, are included in this review. Importantly, the measures included in this review should not be interpreted as diagnostically significant for an anxiety disorder, even generalized anxiety disorder, but should be used to measure the presence of symptoms and to calibrate the severity of general symptoms of anxiety commonly occurring in rheumatic disease. The measures reviewed below include the State Trait Anxiety Index, the Beck Anxiety Inventory, and the anxiety subscale of the Hospital Anxiety and Depression Scale. In this review, the content and structure of each measure is presented (number of items, recall period, response options, presence of translations, and adaptations), the use in rheumatic disease when possible is discussed, and the psychometric properties of each measure, particularly when validated in any of the rheumatic diseases, is detailed. In addition, information regarding responsiveness of each measure to longitudinal change is presented, including responsiveness to change in rheumatology when available. Finally, a summary of the strengths and weaknesses specific to rheumatology is presented.

Supported by the NIH (grants 5-K08-MH072724 and 5-P60-AR053308).

Laura J. Julian, PhD: University of California, San Francisco.

Address correspondence to Laura J. Julian, PhD, Department of Medicine, University of California, San Francisco, 3333 California Street, Suite 270, San Francisco, CA 94143-0920. E-mail: Laura.julian@ucsf.edu.

Submitted for publication January 24, 2011; accepted in revised form July 9, 2011.

THE STATE-TRAIT ANXIETY INVENTORY (STAI)

Description

Purpose. To measure via self-report the presence and severity of current symptoms of anxiety and a generalized propensity to be anxious. Versions of this measure are available for both adults and children.

Content. There are 2 subscales within this measure. First, the State Anxiety Scale (S-Anxiety) evaluates the current state of anxiety, asking how respondents feel "right now," using items that measure subjective feelings of apprehension, tension, nervousness, worry, and activation/arousal of the autonomic nervous system. The Trait Anxiety Scale (T-Anxiety) evaluates relatively stable aspects of "anxiety proneness," including general states of calmness, confidence, and security.

Number of items. The STAI has 40 items, 20 items allocated to each of the S-Anxiety and T-Anxiety subscales. There is also a STAI for children (STAIC) with the same number of items. Short versions of the scales have been developed independently (2–4).

Response options/scale. Responses for the S-Anxiety scale assess intensity of current feelings "at this moment": 1) not at all, 2) somewhat, 3) moderately so, and 4) very much so. Responses for the T-Anxiety scale assess frequency of feelings "in general": 1) almost never, 2) sometimes, 3) often, and 4) almost always.

Examples of use. First published in 1970 with the original STAI-X, the STAI was revised in 1983 (STAI-Y) and has been used extensively in a number of chronic medical conditions including rheumatic conditions such as rheumatoid arthritis (5), systemic lupus erythematosus (6), fibromyalgia, and other musculoskeletal conditions (7).

Practical Application

How to obtain. The STAI can be obtained from the publisher, Mind Garden, 855 Oak Grove Avenue, Suite 215, Menlo Park, CA 94025 (URL: http://www.mindgarden.com/index.htm.) Description of the shortened S-Anxiety scale has been published (2–4), and used in rheumatic disease (rheumatoid arthritis) (8).

Method of administration. Paper and pencil administration. This is a self-report questionnaire that can be administered in an individual format. Specific instructions are provided for each of the S-Anxiety and T-Anxiety subscales.

Scoring. Item scores are added to obtain subtest total scores. Scoring should be reversed for anxiety-absent items (19 items of the total 40). Mind Garden has a service available to administer and score, and there is a web-based interface available through http://www.mindgarden.com/index.htm.

Score interpretation. Range of scores for each subtest is 20–80, the higher score indicating greater anxiety. A cut point of 39–40 has been suggested to detect clinically significant symptoms for the S-Anxiety scale (9,10); however, other studies have suggested a higher cut score of 54–55 for older adults (11). Normative values are available in the manual (12) for adults, college students, and psychiatric samples. To this author's knowledge, no cut scores have been validated for rheumatic disease populations.

Respondent burden. For adults, this measure requires ~10 minutes to complete.

Translations/adaptations. The STAI has been translated and adapted in 48 languages.

Psychometric Information

Reliability. Test–retest reliability coefficients on initial development (12) ranged from 0.31 to 0.86, with intervals ranging from 1 hour to 104 days. Not surprisingly, since the S-Anxiety scale tends to detect transitory states, test–retest coefficients were lower for the S-Anxiety as compared to the T-Anxiety. Internal consistency alpha coefficients were quite high ranging from 0.86 for high school students to 0.95 for military recruits (12).

Validity. During test development, more than 10,000 adults and adolescents were tested. To optimize content validity, most items were selected from other anxiety measures on the basis of strong associations with the Taylor Manifest Anxiety Scale (13) and Cattell and Scheier's Anxiety Scale Questionnaire (14); overall correlations between the STAI and these 2 measures were 0.73 and 0.85, respectively. In general, construct validity (15) of the STAI was somewhat limited in discriminating anxiety from depression, with some studies observing higher correlations between the T-Anxiety scale and measures of depression, as compared to other measures of anxiety (5,16). S-Anxiety validity was originally derived from testing in situations characterized by high state stress including classroom examinations, military training programs, etc. Like other measures of anxiety, the STAI is also highly correlated with depression and, in some studies, the STAI did not differentiate anxious from depressed patients (17). Similarly, while the STAI has not been formally validated in rheumatic disease, studies in rheumatology have similarly observed very high correlations among the STAI and measures of depression (e.g., r = 0.83) (5). In some populations (elderly), the STAI has shown poor discriminant validity and did not differentiate persons with and without anxiety disorders (16).

Ability to detect change. The intent of the T-anxiety scale is to characterize anxiety "proneness" as a longstanding trait or characteristic, and as such, the T-Anxiety is less responsive to change as compared to the S-Anxiety.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The STAI is among the most widely researched and widely used measures of general anxiety, and is available in many different languages. Many use the STAI in rheumatologic conditions. This measure is relatively brief to administer and does not require costly or time consuming scoring or interpretation procedures. Therefore, this measure lends itself well to general use in research in the rheumatology clinic and comparisons with other healthy, psychiatric, and medical populations.

Caveats and cautions. Limitations include the limited availability of validation data specific to rheumatic disease. Additionally, there exists relatively poor validity of the scale, particularly the T-Anxiety subscale for differentiation anxious from depressed states. Further, because the intent of the T-Anxiety scale is to characterize a longstanding trait, clinicians and researchers should be mindful of this if seeking scales to detect change over a relatively short period of time. In general, for these purposes, many have opted to solely use the S-Anxiety subscale for the detection of longitudinal change.

BECK ANXIETY INVENTORY (BAI)

Description

Purpose. The BAI is a brief measure of anxiety with a focus on somatic symptoms of anxiety that was developed as a measure adept at discriminating between anxiety and depression (18).

Content. The BAI is administered via self-report and includes assessment of symptoms such as nervousness, dizziness, inability to relax, etc.

Number of items. The BAI has a total of 21 items.

Response options/scale. Respondents indicate how much they have been bothered by each symptom over the past week. Responses are rated on a 4-point Likert scale and range from 0 (not at all) to 3 (severely).

Examples of use. The BAI is used in efforts to obtain a purer measure of anxiety that is relatively independent of depression. Increasing use of this measure has been observed in a number of rheumatic conditions including fibromyalgia (19) and arthritis (20).

Practical Application

How to obtain. The BAI is not in the public domain, but is a copyrighted measure by the developer, Dr. Aaron T. Beck. The measure can be purchased from Pearson Assessment at www.pearsonassessments.com.

Method of administration. Paper and pencil administered. This is a self-report or interviewer administered questionnaire that can be administered in an individual format.

Score interpretation. Scoring is easily accomplished by summing scores for items. The total score ranges from 0–63. The following guidelines are recommended for the interpretation of scores: 0–9, normal or no anxiety; 10–18, mild to moderate anxiety; 19–29, moderate to severe anxiety; and 30–63, severe anxiety. To this author's knowledge, no published cut scores are available for rheumatologic populations.

Respondent burden. For adults, this measure requires ~5–10 minutes to complete.

Translations/adaptations. The BAI is distributed by Pearson Assessments into Spanish and English. A computer-administered version has been developed by Steer and colleagues (21). The BAI has also been translated into French, German, African languages (e.g., Xhosa), Norwegian, and other languages.

Psychometric Information

Validity. Construct validity studies show good convergence of the BAI with other measures of anxiety including the Hamilton Anxiety Rating Scale (r = 0.51), the STAI (r = 0.47–0.58), and the anxiety scale of the Symptom Checklist-90 (r = 0.81) (22). Although the BAI appears to be less correlated with depression scales than the STAI, correlations with depression scales remain substantial (e.g., correlation with Beck Depression Inventory r = 0.61). While to this author's knowledge, the BAI has not been validated in rheumatology populations, studies among other populations with medical comorbidities (e.g., older adults) suggest that due to the emphasis on somatic symptoms, the BAI did not perform similarly to younger populations (yielded somatic factors in older adults), and therefore the discriminant validity may be less robust than in younger or healthy populations (23).

Reliability. Internal consistency is high with Cronbach's alphas ranging from 0.90 to 0.94 and has been tested in large samples of psychiatric patients, college students, and community-dwelling adults (24–26). Test–retest coefficients are reasonable and range from 0.62 (7-week interval) to 0.93 (1-week interval).

Ability to detect change. The BAI has been demonstrated to be responsive to change over time both on psychiatric populations (27) and in medical populations (28). One study tested the BAI longitudinally over the course of a treatment trial (duloxetine) for the treatment of fibromyalgia and did not show a significant BAI change over time; however, it is important to note that anxiety was not the targeted outcome of this study (19).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BAI is a relatively brief, easily administered, and easily scored measure of anxiety. It has sound psychometric properties and has demonstrated sensitivity to change. This measure has increasing use in a number of rheumatic conditions including fibromyalgia (19) and arthritis (20).

Caveats and cautions. The primary limitations for the BAI are the relatively limited scope of symptoms evaluated and the lack of validation studies specific to rheumatology populations. The BAI was developed in an attempt to reduce overlap with depressive symptoms, and as a result tends to focus more exclusively on somatic (e.g., heart racing, dizziness) symptoms. In medical conditions, these symptoms have the propensity to overlap with some physical aspects of medical conditions and, therefore, cautious interpretation would be warranted. The BAI does not assess other primary symptoms of anxiety, most notably worry and other cognitive aspects of anxiety. In summary, for rheumatology, unless accompanied by other measures that include cognitive (ruminative) aspects of anxiety, the BAI may provide a limited assessment of anxiety.

HOSPITAL ANXIETY AND DEPRESSION SCALE-ANXIETY (HADS-A)

Description

Purpose. The HADS (29) depression component is reviewed elsewhere in this special issue. In general the HADS-A was developed as a brief measure of generalized symptoms of anxiety and fear. The purpose of the HADS was to screen for clinically significant anxiety and depressive symptoms in medically ill patients.

Content. The HADS-A includes specific items that assess generalized anxiety including tension, worry, fear, panic, difficulties in relaxing, and restlessness.

Number of items. The HADS-A has 7 items.

Recall period/response items. Respondents indicate how they currently feel. Responses are rated on a 4-point Likert scale and range from 0 to 3. Anchor points for the Likert items vary depending on the item (e.g., "I can sit still and feel relaxed" scores as 0 for definitely to 3 for not at all; and "I get sudden feelings of panic" scores as 0 for not at all to 3 for very much indeed).

Examples of use. This measure evaluates common dimensions of anxiety. This measure can be used to detect and quantify magnitude of symptoms of anxiety, but like other measures is not adequately descriptive to detect specific anxiety disorders. The target population is general medical outpatients age 16 to 65.

Practical Application

How to obtain. The HADS is copyrighted and available from: Nfer Nelson, The Chiswick Centre, 414 Chiswick

High Road, London W4 5TF United Kingdom. URL: www.nfer-nelson.co.uk.

Method of administration. Paper and pencil administered. This is an individually administered questionnaire and can be given via self-report or by interviewer.

Score interpretation. Scoring is easily accomplished by summing scores for items, with special attention to reversed items. The total score for the HADS-A can range from 0 to 21. The following guidelines are recommended for the interpretation of scores: 0–7 for normal or no anxiety, 8–10 for mild anxiety, 11–14 for moderate anxiety, and 12–21 for severe anxiety. In some rheumatologic conditions, a cut score for the HADS-A of 9 was recommended as useful for a diagnosis of an anxiety disorder (30).

Respondent burden. For adults, this measure typically requires <5 minutes to complete.

Translations/adaptations. Translations are available in Arabic, Chinese, Dutch, French, German, Hebrew, Japanese, Italian, Spanish, and Urdu.

Psychometric Information

Validity. The majority of psychometric studies observed a 2-factor solution, supporting the use of the anxiety subscale as a "stand alone" measure (11 of 19 studies in a recent review of this measure; however a few studies did find more than 2 factors (see review by Bjelland et al [31]). Using a cut score of 8 overall provided sensitivities and specificities at ~80% and reaching 90% in a community cohort for the HADS-A for detecting anxiety disorders (31). In primary care populations, cut scores of ≥9 for the HADS-A yielded moderate sensitivity (0.66) and high specificity (0.93) (31). An additional study in the elderly yielded high misclassification rates and suggested that the HADS-A possessed limited sensitivity and specificity to detect anxiety disorders in this population (32). One study comparing the HADS to diagnoses of anxiety and depression in a cohort of patients with osteoarthritis observed greater concordance among the HADS-A and diagnoses of anxiety compared to the concordance among the HADS depression scale and diagnoses of depressive disorders (30). In this study, the HADS-A had a sensitivity and specificity of 88% and 81%, respectively, for a diagnosis of an anxiety disorder (33,34). Overall, concurrent validity of the HADS was deemed "good" to "very good" in a comprehensive review (31), with comparable sensitivity and specificity of longer scales including the General Health Questionnaire, the STAI, and the Symptom Checklist-90 anxiety scales.

Reliability. Internal consistency is high for the anxiety component with Cronbach's alphas ranging from 0.84–0.90 and has been tested in large samples of community dwelling adults, psychiatric samples, and medical samples (33,35,36).

Ability to detect change. There is some evidence, including through the use of change reliability indices, that the HADS-A is sensitive to change (37). In particular, the HADS-A has been found to be responsive to change longitudinally in ankylosing spondylitis (38), and other arthritis populations (39).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The HADS-A is a very brief, easy to use screening measure to detect the presence of clinically significant symptoms of anxiety designed for use in medical populations. This measure is widely used and easily obtained. The splitting of the subscales (anxiety and depression) is a commonly used practice, and there are data supporting the use of the HADS-A as a stand-alone measure of general anxiety. The HADS has been widely used in rheumatologic populations including Sjögren's syndrome (40), ankylosing spondylitis (38), various forms of arthritis (39,41,42), and systemic lupus erythematosus (43).

Caveats and cautions. Weaknesses include some evidence of reduced validity in some populations, particularly in the elderly. Like other measures reviewed here, this measure does not adequately detect the presence of specific anxiety disorders, but rather provides some evidence towards generalized anxiety symptoms.

DISCUSSION

Three measures were reviewed above: the STAI, the BAI, and the HADS-A. These 3 measures were selected for review based on the previous use in rheumatology, sound psychometric properties, and detection of generalized symptoms of anxiety. As mentioned above, measures targeted towards the assessment of specific anxiety disorders including other DSM-IV anxiety disorders (including posttraumatic stress disorder, obsessive-compulsive disorder, etc.) are not included in this review. While assessment of some of these features may be beneficial in rheumatology, for example, some studies in other populations have observed posttraumatic stress type reactions to receiving specific medical diagnoses (44,45), these instances are more unique considerations and, therefore, such measures are not included in this review.

It becomes evident, based on the brevity of this review, that few stand-alone measures of anxiety are currently used in rheumatology. Reasons for the decreased emphasis on the assessment of anxiety in these populations may be multifaceted and include a relative increased emphasis on depression in comparison to anxiety, use of larger scale measures detecting a range of features related to psychological distress (e.g., Symptom Checklist-90), or an underappreciation of the prevalence and severity of anxiety in many rheumatic conditions. Moving forward, it may be warranted to explore these factors more fully and determine if the current measures in use are adequately detecting the presence and severity of symptoms of anxiety that are important to patients or that need to be addressed in the course of medical care. Nonetheless, based on this review, there currently exist measures that have good psychometric properties and adequate responsiveness to change that would warrant use in rheumatology.

AUTHOR CONTRIBUTIONS

Dr. Julian drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Washington (DC): American Psychiatric Association; 1994.
2. Tluczek A, Henriques JB, Brown RL. Support for the reliability and validity of a six-item state anxiety scale derived from the State-Trait Anxiety Inventory. J Nurs Meas 2009;17:19–28.
3. Kaipper MB, Chachamovich E, Hidalgo MP, Torres IL, Caumo W. Evaluation of the structure of Brazilian State-Trait Anxiety Inventory using a Rasch psychometric approach. J Psychosom Res 2010;68:223–33.
4. Chlan L, Savik K, Weinert C. Development of a shortened state anxiety scale from the Spielberger State-Trait Anxiety Inventory (STAI) for patients receiving mechanical ventilatory support. J Nurs Meas 2003; 11:283–93.
5. VanDyke MM, Parker JC, Smarr KL, Hewett JE, Johnson GE, Slaughter JR, et al. Anxiety in rheumatoid arthritis. Arthritis Rheum 2004;51: 408–12.
6. Ward MM, Marx AS, Barry NN. Psychological distress and changes in the activity of systemic lupus erythematosus. Rheumatology (Oxford) 2002;41:184–8.
7. White KP, Nielson WR, Harth M, Ostbye T, Speechley M. Chronic widespread musculoskeletal pain with or without fibromyalgia: psychological distress in a representative community adult sample. J Rheumatol 2002;29:588–94.
8. Wong M, Mulherin D. The influence of medication beliefs and other psychosocial factors on early discontinuation of disease-modifying anti-rheumatic drugs. Musculoskeletal Care 2007;5:148–59.
9. Knight RG, Waal-Manning HJ, Spears GF. Some norms and reliability data for the State-Trait Anxiety Inventory and the Zung Self-Rating Depression scale. Br J Clin Psychol 1983;22 Pt 4:245–9.
10. Addolorato G, Ancona C, Capristo E, Graziosetto R, Di Rienzo L, Maurizi M, et al. State and trait anxiety in women affected by allergic and vasomotor rhinitis. J Psychosom Res 1999;46:283–9.
11. Kvaal K, Ulstein I, Nordhus IH, Engedal K. The Spielberger State-Trait Anxiety Inventory (STAI): the state scale in detecting mental disorders in geriatric patients. Int J Geriatr Psychiatry 2005;20:629–34.
12. (Primary reference) Spielberger C. Manual for the State-Trait Anxiety Inventory (rev. ed.). Palo Alto (CA): Consulting Psychologists Press; 1983.
13. Taylor JA. A personality scale of manifest anxiety. J Abnorm Psychol 1953;48:285–90.
14. Cattell RB, Sheier IH. Handbook for the IPAT Anxiety Scale. 2nd ed. Champaign (IL): Institute for Personality and Ability Testing; 1963.
15. Spielberger CD. State-Trait Anxiety Inventory. Palo Alto (CA): Consulting Psychologists Press; 1983.
16. Kabacoff RI, Segal DL, Hersen M, Van Hasselt VB. Psychometric properties and diagnostic utility of the Beck Anxiety Inventory and the State-Trait Anxiety Inventory with older adult psychiatric outpatients. J Anxiety Disord 1997;11:33–47.
17. Kennedy BL, Schwab JJ, Morris RL, Beldia G. Assessment of state and trait anxiety in subjects with anxiety and depressive disorders. Psychiatric Quarterly 2001;72:263–76.
18. (Primary reference) Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol 1988;56:893–7.
19. Arnold LM, Clauw D, Wang F, Ahl J, Gaynor PJ, Wohlreich MM. Flexible dosed duloxetine in the treatment of fibromyalgia: a randomized, double-blind, placebo-controlled trial. J Rheumatol 2010;37: 2578–86.
20. Scopaz KA, Piva SR, Wisniewski S, Fitzgerald GK. Relationships of fear, anxiety, and depression with physical function in patients with knee osteoarthritis. Arch Phys Med Rehabil 2009;90:1866–73.
21. Steer RA, Rissmiller DJ, Ranieri WF. Structure of the computer-assisted Beck Anxiety Inventory with psychiatric inpatients. J Pers Assess 1993; 60:532–42.
22. Beck AT, Steer RA. Relationship between the Beck Anxiety Inventory and the Hamilton Anxiety Rating Scale with anxious outpatients. J Anx Disord 1991;5:213–23.
23. Morin C, Landreville P, Colecchi C, McDonald K, Stone J, Ling W. The Beck Anxiety Inventory: psychometric properties with older adults. J Clin Geropsychol 1999;5:19–29.
24. Fydrich T, Dowdall D, Chambless DL. Reliability and validity of the Beck Anxiety Inventory. J Anx Disord 1993;6:55–61.
25. Creamer M, Foran J, Bell R. The Beck Anxiety Inventory in a non-clinical sample. Behav Res Ther 1995;33:477–85.
26. Osman A, Barrios FX, Aukes D, Osman JR. The Beck Anxiety Inventory: psychometric properties in a community population. J Psychopath Behav Assess 1993;15:287–97.
27. Brown GK, Beck AT, Newman CF. A comparison of focused and standard cognitive therapy for panic disorder. J Anx Disord 1997;11:329–45.
28. Lee YW, Park EJ, Kwon IH, Kim KH, Kim KJ. Impact of psoriasis on quality of life: relationship between clinical response to therapy and change in health-related quality of life. Ann Dermatol 2010;22:389–96.
29. (Primary reference) Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361–70.
30. Axford J, Butt A, Heron C, et al. Prevalence of anxiety and depression in osteoarthritis: use of the Hospital Anxiety and Depression Scale as a screening tool. Clin Rheumatol 2010;29:1277–83.
31. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. J Psychosom Res 2002;52:69–77.
32. Davies KN, Burn WK, McKenzie FR, Brothwell JA, Wattis JP. Evaluation of the hospital anxiety and depression scale as a screening instrument in geriatric medical inpatients. Int J Geriatr Psychiatry 1993;8: 165–9.
33. Lisspers J, Nygren A, Soderman E. Hospital Anxiety and Depression Scale (HAD): some psychometric data for a Swedish sample. Acta Psychiatr Scand 1997;96:281–6.
34. Lepine JP, Godchau M, Brun P, Lemperiere T. Evaluation of anxiety and depression among patients hospitalized on an internal medicine service. Ann Med Psychol (Paris) 1985;143:175–89. In French.
35. Bedford A, de Pauw K, Grant I. The structure of the Hospital Anxiety and Depression scale (HAD): an appraisal with normal, psychiatric, and medical patient subjects. Pers Indiv Differ 1997;23:473–8.
36. Dagnan D, Chadwick P, Trower P. Psychometric properties of the Hospital Anxiety and Depression Scale with a population of members of a depression self-help group. Br J Med Psychol 2000;73:129–37.
37. Hinz A, Zweynert U, Kittel J, Igl W, Schwarz R. Measurement of change with the Hospital Anxiety and Depression Scale (HADS): sensitivity and reliability of change. Psychother Psychosom Med Psychol 2009; 59:394–400. In German.
38. Ertenli I, Ozer S, Kiraz S, Apras SB, Akdogan A, Karadag O, et al. Infliximab, a TNF-alpha antagonist treatment in patients with ankylosing spondylitis: the impact on depression, anxiety and quality of life level. Rheumatol Int 2010:1–8.
39. Buszewicz M, Rait G, Griffin M, Nazareth I, Patel A, Atkinson A, et al. Self management of arthritis in primary care: randomised controlled trial. BMJ 2006;333:879.
40. Valtysdottir ST, Gudbjornsson B, Lindqvist U, Hallgren R, Hetta J. Anxiety and depression in patients with primary Sjögren's syndrome. J Rheumatol 2000;27:165–9.
41. Murphy H, Dickens C, Creed F, Bernstein R. Depression, illness perception and coping in rheumatoid arthritis. J Psychosom Res 1999;46: 155–64.
42. Cauli A, Gladman DD, Mathieu A, Olivieri I, Porru G, Tak PP, et al. Patient global assessment in psoriatic arthritis: a multicenter GRAPPA and OMERACT study. J Rheumatol 2011;38:898–903.
43. Mak A, Tang C, Chan M-F, Cheak A, Ho R. Damage accrual, cumulative glucocorticoid dose and depression predict anxiety in patients with systemic lupus erythematosus. Clin Rheumatol 2011;30:795–803.
44. Roberge M, Dupuis G, Marchland A. Post-traumatic stress disorder following myocardial infarction: prevalence and risk factors. Can J Cardiol 2010;26:e170–5.
45. Elklit A, Blum A. Psychological adjustment one year after the diagnosis of breast cancer: a prototype study of delayed post-traumatic stress disorder. Br J Clin Psychol. E-pub ahead of print.

Summary Table for Measures of Anxiety*

								Psychometric properties				
Scale	**Content**	**No. items**	**Response format**	**Administration**	**Time, minutes**	**Primary scale outputs**	**Validated populations**	**Reliability**	**Validity**	**Responsiveness**	**Strengths**	**Cautions**
STAI	Current "state" anxiety; pervasive "trait" anxiety	40, 20 per scale	4-point Likert scale State: symptom intensity Trait: symptom frequency	Self-report (individual or group)	10	Severity of state/trait anxiety	General and psychiatric	Good	Moderate	State anxiety more responsive to change than trait anxiety subscale	Widely used Available in many languages Detection of pervasive anxiety "proneness" and current symptoms	Trait scale measures longstanding traits and therefore is less sensitive to change over a short period of time
BAI	Symptoms of anxiety with a focus on somatic symptoms	21	4-point Likert scale (0 = not at all; 3 = severely)	Self report or interviewer administered	5–10	Total anxiety score	General and psychiatric	Good	Moderate	Established responsiveness to change in psychiatric and medical populations	Brief Sound psychometrics	Relatively narrow scope of symptom assessment with focus on somatic symptoms
HADS-A	Generalized symptoms of anxiety and fear	7	Total anxiety 4-point Likert scale (0 = symptom absent; 3 = symptom present)	Self-report	≤5	Total anxiety score	Medical including arthritis	Excellent	Good	Sensitive to change	Brief Widely used in rheumatology Strongest psychometric properties	Not appropriate to detect specific anxiety disorders May have reduced validity in some populations (e.g., elderly)

* STAI = State-Trait Anxiety Inventory; BAI = Beck Anxiety Inventory; HADS-A = Hospital Anxiety and Depression Scale-Anxiety.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S486–S490
DOI 10.1002/acr.20581
© 2011, American College of Rheumatology

COMMENTARY

The Future of Measuring Patient-Reported Outcomes in Rheumatology

Patient-Reported Outcomes Measurement Information System (PROMIS)

DINESH KHANNA,[1] ESWAR KRISHNAN,[2] ESI MORGAN DEWITT,[3] PUJA P. KHANNA,[1]
BRENNAN SPIEGEL,[4] AND RON D. HAYS[4]

INTRODUCTION

The National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS; trademarked by the National Institutes of Health) Roadmap initiative (available at www.nihpromis.org) is a cooperative research program designed to develop, evaluate, and standardize item banks to measure patient-reported outcomes across different medical conditions as well as the US population (1). The goal of PROMIS is to develop reliable and valid item banks using item response theory (IRT) that can be administered in a variety of formats including short forms and computerized adaptive tests (CAT) (1–3). IRT is often referred to as "modern psychometric theory," in contrast to "classic test theory," or CTT. The basic idea behind both IRT and CTT is that there is some latent construct, or "trait," underlying an illness experience. This construct cannot be directly measured, but can be indirectly measured by creating items that are scaled and scored. For example, fatigue, pain, disability, or even "happiness" are latent constructs, i.e., subjective feelings—we cannot take a picture, snap a radiograph to view them, or run a blood test to check for them. However, we know they exist. People can experience more or less of these constructs; therefore, it is helpful to try to translate that experience into several levels represented by scores. IRT models the associations between items and the latent construct. Specifically, IRT models describe relationships between a respondent's underlying level on a construct and the probability of particular item responses.

Tests developed with CTT (such as the Health Assessment Questionnaire disability index [HAQ DI] [4], or the Scleroderma Gastrointestinal Tract instrument [5]) require administering all items, even though only some are appropriate for the person's trait level. Some items are too high for those with low trait levels (e.g., "can you walk 100 yards" to a patient in a wheelchair) or too low for those with high trait levels (e.g., "can you get up from the chair?" to a runner). In contrast, IRT methods make it possible to estimate person trait levels with any subset of items appropriate for the person's trait levels in an item pool. As such, any set of items from the pool could be administered as a fixed form or, for greatest efficiency, administered as a CAT. CAT is an approach to administering the subset of items in an item bank that are most informative for measuring the health construct in order to achieve a target standard error of measurement. A good item bank will have items that represent a range of content and difficulty, provide high levels of information, and have items that perform equivalently in different subgroups of the target population.

HOW DOES CAT WORK?

Without prior information, the first item administered in a CAT is typically one of medium trait level. For example, "In the past 7 days I was grouchy" with multilevel responses ranging from "never" to "always." After each response, the person's trait level and associated standard error are estimated. The next item administered to someone not endorsing the first item is an easier item. If the

Dr. D. Khanna's work was supported by the NIH (grants U01-AR-057936A and K23-AR-053858-05). Dr. Krishnan's work was supported by the NIH (grant 2U01-AR-052158). Dr. DeWitt's work was supported by the NIH (grant U01-AR-057940). Dr. P. Khanna's work was supported by the NIH (grant U01-AR-057936A) and a Career Bridge Funding Award from the American College of Rheumatology. Dr. Spiegel's work was supported by the NIH (grants U01-AR-057936A, AR-052177). Dr. Hays' work was supported by the NIH (grants U01-AR-057936A, P30-AG-021684, and 2P20-MD-000182).

[1]Dinesh Khanna, MD, MS, Puja P. Khanna, MD, MPH: University of Michigan School of Medicine, Ann Arbor; [2]Eswar Krishnan, MD: Stanford University, Palo Alto, California; [3]Esi Morgan DeWitt, MD, MSCE: Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; [4]Brennan Spiegel, MD, MSHS, Ron D. Hays, PhD: University of California, Los Angeles.

Address correspondence to Dinesh Khanna, MD, MSc, Associate Professor of Medicine and Director, University of Michigan Scleroderma Program, 24 Frank Lloyd Wright Drive, PO Box 481, Ann Arbor, MI 48106. E-mail: khannad@med.umich.edu.

Submitted for publication May 30, 2011; accepted in revised form August 4, 2011.

person endorses the first item, the next item administered is a harder item. CAT is terminated when the standard error falls below an acceptable value. This provides an estimate of one's score with the minimal number of questions and no loss of measurement precision. In addition, scores from different studies using different items can be compared using a common scale. IRT models estimate the underlying scale score (theta) from the items. All items are calibrated on the same metric and independently and collectively provide an estimate of theta. Hence, it is possible to estimate the score using any subset of items and to estimate the standard error of the estimated score. This allows assessment of health outcomes across patients with differing medical conditions (such as comparing scores of someone with arthritis to someone with heart disease) at various degrees of physical and other impairments, both at the lowest and highest ends of trait levels.

PROMIS IN RHEUMATOLOGY

The Life Story of PROMIS Tools

Since the beginning of PROMIS in 2004, much progress has been made in developing measures of self-reported health within a domain hierarchy (Figure 1). Physical functioning, fatigue, pain, emotional distress, and social health were the core domains of interest. While all these domains are relevant to rheumatic diseases, the physical health domain encompassed most of the traditionally important outcomes in rheumatology, such as physical function, pain, and fatigue.

In PROMIS, the term physical function was preferred over the term disability and represented the ability to perform activities of daily living including instrumental activities (e.g., shopping) (6). The PROMIS physical function item bank containing 124 new items was developed from 1,865 available items culled from 160 English language questionnaires. In addition to administering the item bank using CAT, PROMIS has developed several static short forms including: 1) a 20-item PROMIS HAQ, which corresponds to the HAQ DI, 2) a PROMIS 10-item static, or short form with items selected as the "best" from the physical function items, and 3) a PROMIS 20-item static form also selected from the "best" PROMIS items. PROMIS HAQ differs from the HAQ DI by deleting the 1-week time frame and increasing the response option set from the original 4 choices to 5 by adding "with a little difficulty." Measurement properties of different PROMIS item banks (PROMIS HAQ, PROMIS 10-item short form, PROMIS 20-item short form, and 10-item PROMIS CAT) were compared to the HAQ DI and physical functioning 10-item scale (PF-10) of the Short Form 36 in 378 patients with rheumatoid arthritis, osteoarthritis, and normal aging cohorts (7). PF-10 provided the least content information followed by HAQ DI, which was better for patients with physical disability (SD less than or equal to −1) but performed poorly for the average population (Figure 2).

PROMIS items (10 or 20 items) performed better than PF-10 and HAQ DI. The PROMIS CAT outperformed all the static items (Figure 2). The CAT maintained acceptable performance in populations whose physical function is 1.5 SDs better than the population norm. This has implications for our patients because as better treatments become available for rheumatic diseases we are likely to observe healthier cohorts of patients with arthritis. Thus, accurate assessment of those in the positive health range of physical functioning becomes increasingly important.

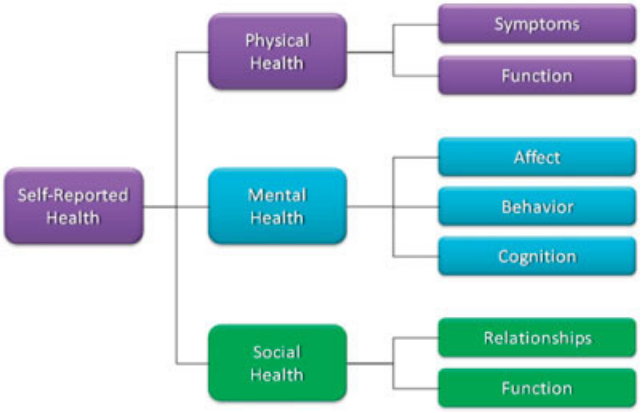


Figure 1. Patient-Reported Outcomes Measurement Information System (PROMIS) domain hierarchy. Color figure can be viewed in the online issue, which is available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1529-0131.

What PROMIS Means for Rheumatology

Physical function, global health assessment, and fatigue are important constructs in rheumatic diseases, in both adults and pediatrics. The availability of PROMIS tools will also catalyze research on the less well-studied impact of rheumatic diseases in all health domains. In the next sections, we discuss the advantages of PROMIS, its current use in rheumatology, and the future of PROMIS in rheumatology.

Advantages of PROMIS over traditional instruments. PROMIS employes a uniform qualitative process with detailed systematic review, focus groups, cognitive interviews, and translatability for each item bank. PROMIS has devoted substantive resources to ensuring that outcome measures are understood and usable by diverse populations. Items are written at a grade school level and tested for comprehensibility among low-literacy populations. All items are reviewed and modified as needed for their translatability. To enhance inclusiveness, PROMIS informatics assessment tools are rendered accessible to populations with sensory limitations and others requiring assistive technology. Lastly, PROMIS measures are grounded in a life course perspective, as it is the group's ultimate goal to produce single metrics for the same domain across the full lifespan (i.e., PROMIS is linking measures developed for children with those developed for adults).

PROMIS instruments have been found to have better precision than existing measures; a quality that may lead to reduction in sample size in clinical studies (6). The severity of patient-reported outcomes in rheumatic diseases can be compared head-to-head with other chronic conditions such as heart failure. It is possible to "customize" the set of items by selecting a set of items that is matched to the severity level of the target population.

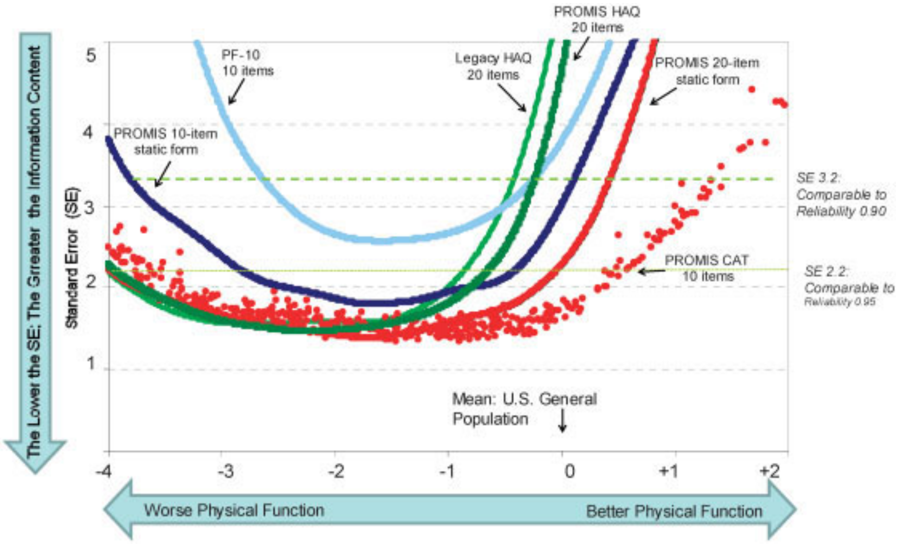


Figure 2. Comparison of information content of 6 physical function instruments: Health Assessment Questionnaire disability index (HAQ DI), Patient-Reported Outcomes Measurement Information System (PROMIS) HAQ, PROMIS 10-item short form, PROMIS 20-item short form, physical functioning 10-item score (PF-10; from Short Form 36), and 10-item PROMIS computer (CAT). Instruments with greater information content have SE curves that are lower and have a greater SD range at a reliability >0.95. More items are better than fewer, item response theory (IRT)-based (PROMIS) is better than non–IRT-based items, and CAT is better than static. Adapted with permission from ref. 7.

PROMIS items are currently available at no cost, enabling freer exchange of information and data, stimulating outcomes research.

Utilization of CAT to administer PROMIS items does require a computer, and that may limit its applicability in a busy clinical practice. Although a person may receive different set of items from an item pool at each visit, users can track which items were administered in the CAT and track theta scores over time.

Current PROMIS item banks and their validation in rheumatology. *PROMIS item banks for adult patients.* PROMIS item banks developed for adults (including anger, anxiety, abilities and general concern, depression, fatigue, pain behavior, pain interference, physical function, positive and negative psychosocial impact of illness, sleep disturbance, sleep impairment, satisfaction with participation in social roles, and satisfaction with participation in discretionary social activities) are available at www.nihpromis.org. Additional short forms have been developed for constructs such as global health, global satisfaction with sex life, etc. All these item banks measure important constructs that are applicable to patients with arthritis and other rheumatologic conditions. As an example, the feasibility of 11 PROMIS item banks was recently assessed in a single-center, observational study in patients with systemic sclerosis (8). The average number of items completed for each CAT-administered item bank ranged from 5 to 8 (69 CAT items per patient), and the average time to complete each CAT-administered item bank ranged from 48 seconds to 1.9 minutes per patient (average time 11.9 minutes/per patient for 11 banks). The time to complete the item banks was not significantly different in patients with physical disabilities (such as hand contractures and digital ulcers).

PROMIS item banks for pediatric patients. PROMIS version 1.0 item banks and short forms developed for children include anger, anxiety, asthma impact, depressive symptoms, fatigue, pain interference, physical function (separate banks for upper extremity and mobility), and peer relationships and are available at www.nihpromis.org. The PROMIS Cooperative Network is currently in the process of evaluating the pediatric version 1.0 item banks in multiple pediatric chronic conditions including juvenile idiopathic arthritis (JIA) and chronic musculoskeletal pain, widespread or regional. Importantly, the process includes a qualitative component including semistructured interviews with children. Longitudinal validation in these pediatric conditions, among others, is underway.

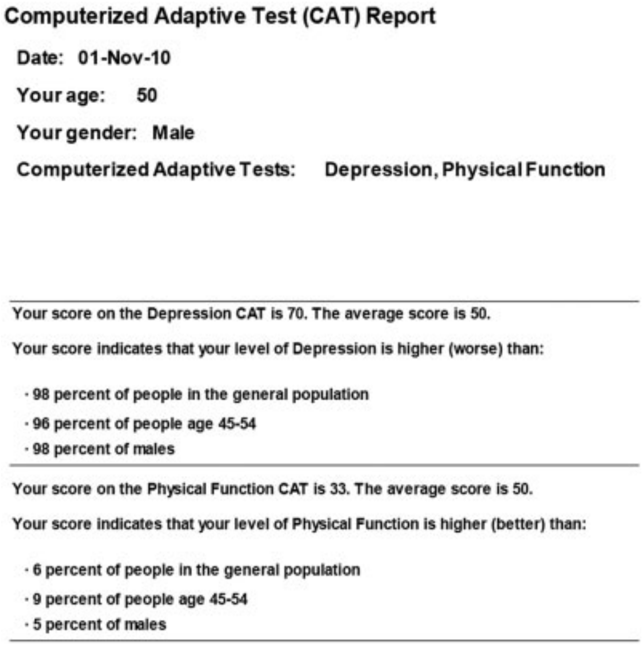
New PROMIS item banks under development. The PROMIS Cooperative Network has increased the focus and energy on development of pediatric item banks with 4 of 12 of the PROMIS II sites (project period 2009–2012) dedicated to work in pediatrics. This includes development of new pediatric item banks to assess pain behavior, pain quality, physical activity, subjective well-being, experience of stress and others, all of which are important in patients with chronic arthritis. Current efforts are also focused on linking adult and pediatric item banks measuring the same construct to allow measurement from childhood through adolescence then transition to adulthood on the same metric. The PROMIS Cooperative Network is also developing new item banks pertinent to chronic diseases. These include development of gastrointestinal symptoms items, self-efficacy for self-management of chronic illness, and others.

Future of PROMIS in rheumatology. The PROMIS mission is to use measurement science to create a state-of-the-art assessment system for self-reported health to advance patent-reported outcome measurement in clinical research and day-to-day practice. Similar to other patient-reported outcomes, this will facilitate the incorporation of the patient's voice into clinical trials and clinical practice. The American College of Rheumatology has endorsed the assessment of functional status in patients with rheumatoid arthritis at least every 12 months. For patients with JIA, it is recommended that functional status and health-related quality of life be assessed at 6 month intervals (9). This exacts new requirements of patient-reported outcome measures, including exceptional ease of use, rapidity of administration, interpretability, and a clear benefit of using the data in patient-provider interactions and care management. Rheumatology is a specialty that is well versed in the use of measures of disability, pain, and other aspects of health-related quality of life. PROMIS offers an opportunity to accelerate uptake and expand the use of patient-reported outcomes from research advocates to all clinicians.

Using PROMIS in Clinical Practice

Being able to administer a choice of fatigue, pain interference, physical function, or depression measures, among many other options, in the waiting room on a Tablet, laptop, personal computer, and potentially a Smart Phone and have instant scoring, calibration to population norms, and be ready to share with the patient at point of care is compelling.

As an example, Figures 3 and 4 show results from a 50-year-old patient with early diffuse systemic sclerosis. This patient was administered CAT item banks for physical function and depression that took approximately 2 minutes to complete. The profile provides his current physical function (1.7 SDs below US general population) and depression status (2 SDs below US population). This information (presented in the form of a PROMIS report in Figure 3 and a graph as shown in Figure 4) can be used for clinical care. This patient was referred for psychological counseling to help him adjust to his newly diagnosed systemic sclerosis and also prescribed physical therapy. The item banks can be administered at each clinic visit to assess change in symptoms from baseline visit. Current work is ongoing to assess the feasibility of incorporating PROMIS item banks in routine clinical practice.



Computerized Adaptive Test (CAT) Report

Date: 01-Nov-10

Your age: 50

Your gender: Male

Computerized Adaptive Tests: Depression, Physical Function

Your score on the Depression CAT is 70. The average score is 50.

Your score indicates that your level of Depression is higher (worse) than:

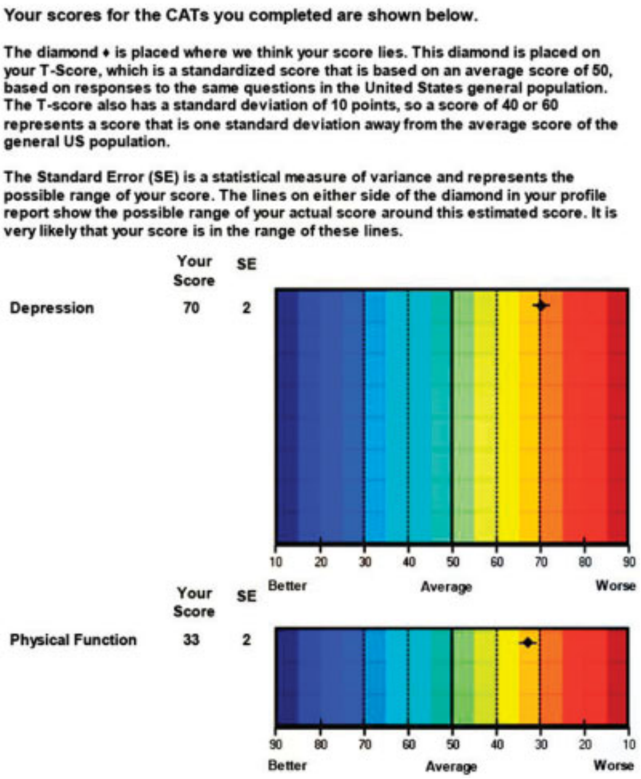
- 98 percent of people in the general population
- 96 percent of people age 45-54
- 98 percent of males

Your score on the Physical Function CAT is 33. The average score is 50.

Your score indicates that your level of Physical Function is higher (better) than:

- 6 percent of people in the general population
- 9 percent of people age 45-54
- 5 percent of males

Figure 3. Computer generated Patient-Reported Outcomes Measurement Information System (PROMIS) report of a 50-year-old man with recently diagnosed systemic sclerosis. The report provides the patient's scores for depression and physical function scales and compares it to the US general population.



Your scores for the CATs you completed are shown below.

The diamond ♦ is placed where we think your score lies. This diamond is placed on your T-Score, which is a standardized score that is based on an average score of 50, based on responses to the same questions in the United States general population. The T-score also has a standard deviation of 10 points, so a score of 40 or 60 represents a score that is one standard deviation away from the average score of the general US population.

The Standard Error (SE) is a statistical measure of variance and represents the possible range of your score. The lines on either side of the diamond in your profile report show the possible range of your actual score around this estimated score. It is very likely that your score is in the range of these lines.

Figure 4. Computer-generated Patient-Reported Outcomes Measurement Information System (PROMIS) graph of a 50-year-old man with recently diagnosed systemic sclerosis. The report depicts the patient's score for depression and physical function scales and compares it to the US general population. Color figure can be viewed in the online issue, which is available at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1529-0131.

In conclusion, PROMIS has developed items banks that are relevant to rheumatology, can be "customized" to a patient or a practice, and are currently freely available. The item banks provide tremendous flexibility for creation of fixed length short forms or CAT administration. This quick assessment can generate a patient report to monitor health over time.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45 Suppl 1:S3–11.
2. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45 Suppl 1:S22–31.
3. Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. Med Care 2007;45 Suppl 1:S32–8.
4. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
5. Khanna D, Hays RD, Maranian P, Seibold JR, Impens A, Mayes MD, et al. Reliability and validity of the University of California, Los Angeles Scleroderma Clinical Trial Consortium Gastrointestinal Tract Instrument. Arthritis Rheum 2009;61:1257–63.
6. Fries JF, Krishnan E, Bruce B. Items, instruments, crosswalks, and PROMIS. J Rheumatol 2009;36:1093–5.
7. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061–6.
8. Khanna D, Maranian P, Rothrock N, Cella D, Gershon R, Khanna PP, et al. Feasibility and evaluation of the construct validity of PROMIS and "Legacy" instruments in an academic scleroderma clinic. Value Health. In press.
9. Lovell DJ, Passo MH, Beukelman T, Bowyer SL, Gottlieb BS, Henrickson M, et al. Measuring process of arthritis care: a proposed set of quality measures for the process of care in juvenile idiopathic arthritis. Arthritis Care Res (Hoboken) 2011;63:10–6.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S371–S382
DOI 10.1002/acr.20635
© 2011, American College of Rheumatology

Measures of Pediatric Function

Child Health Assessment Questionnaire (C-HAQ), Juvenile Arthritis Functional Assessment Scale (JAFAS), Pediatric Outcomes Data Collection Instrument (PODCI), and Activities Scale for Kids (ASK)

SUSAN E. KLEPPER

INTRODUCTION

Children with a rheumatic disease frequently experience impairments in one or more body systems; these may include pain, stiffness, fatigue, muscle weakness, soft tissue contractures, and poor exercise capacity. These impairments may directly limit the child's ability to perform some physical activities or may do so indirectly if the child or parent fears such activities may cause injury or a disease flare. It is essential to understand the impact of rheumatic disease on a child's activities in order to guide intervention and monitor changes in functional abilities over time and with targeted therapies. This is especially important when disease onset occurs at a very young age because the long-term effects of physical limitations can negatively impact the child's quality of life.

The best measure of activities for a particular child or group of children depends on the context of the evaluation, including the physical and social environment. For example, a child may perform a task like standing up from the floor without assistance in a quiet, standardized environment like the clinic, but does not perform the same task during physical education class at school or after a fall in the community. Holsbeeke et al (1) suggest there are 3 related but separate constructs of physical activity that are distinguished by this person-environment interaction. Capacity describes what a person can do in a standardized controlled environment, capability describes what a person can do in his/her daily environment, and performance describes what a person actually does in his/her daily environment.

Four measures of physical activities that have been developed for or are appropriate for use in children with a rheumatic disease will be reviewed. Only the Juvenile Arthritis Functional Assessment Scale (JAFAS) and the Childhood Health Assessment questionnaire (C-HAQ) were developed specifically for children with juvenile arthritis. The C-HAQ, the most frequently used measure of activities in pediatric rheumatology, evaluates a child's capability to perform activities in their daily environment, while the JAFAS measures the child's capacity in the daily environment. Both the Activities Scale for Kids (ASK) and the Pediatric Outcomes Data Collection Instruments (PODCI), although not specifically designed for use in pediatric rheumatic diseases, assess physical function in children with chronic health disorders, including childhood arthritis. The ASK includes 2 versions, one that measures a child's physical capability in his/her daily environment (ASK_c) and one that measures the child's performance of the same activities in their daily environment (ASK_p). The ASK is also the only measure that requires the child to be the respondent because, as Young et al (2) state, "it is the child who is most familiar with his or her own abilities or limitations in each setting." Finally, the PODCI, the most comprehensive of the 4 instruments, measures capability primarily, and includes a pediatric version to be completed by a parent and an adolescent version that can be completed by the parent, child, or both. Each of these 4 measures includes activities that are necessary and important to children across a wide age range. The use of 2 or more of these measures in combination may provide clinicians with the best understanding of a child's typical activities and participation in age-appropriate settings.

CHILDHOOD HEALTH ASSESSMENT QUESTIONNAIRE (C-HAQ)

Description

Purpose. To assess health status and physical function in children with juvenile arthritis. The populations are children, ages 1–19 years with juvenile arthritis (3). It is also validated for use in children with juvenile der-

Susan E. Klepper, PhD, PT: Columbia University, New York, New York.

Dr. Klepper has received honoraria (less than $10,000) from The Association of Rheumatology Health Professionals.

Address correspondence to Susan E. Klepper, PhD, PT, Assistant Professor of Clinical Physical Therapy, Columbia University, 710 West 168th St., 8th Floor, New York, NY 10032. E-mail: sek44@columbia.edu.

Submitted for publication January 31, 2011; accepted in revised form May 9, 2011.

matomyositis (4,5) and systemic lupus erythematosus (6), and used for children with chronic musculoskeletal pain syndromes (7), and spina bifida (8). The original publication was by Singh et al in 1994 (3). Lam et al (9) proposed several revised versions that included 8 additional items to assess the child's functional strengths as well as deficits.

Content. Includes the International Classification of Functioning, Disability, and Health components of body function (sensation of pain) and activities and participation (basic and instrumental activities of daily living [ADLs] considered to be important, and often difficult for children with arthritis), as well as a measure of overall health status. Revised versions contain 8 additional more physically challenging activities.

Number of items. The disability index (DI) in the original C-HAQ includes 30 items grouped into 8 domains of physical function, dressing and grooming (4 items), arising (2 items), eating (3 items), walking (2 items), hygiene (5 items), reach (4 items), grip (5 items), and activities (5 items). The DI in revised versions (C-HAQ$_{38}$) includes 8 additional, more physically-challenging items (9). The discomfort index (pain) and health status index (overall health status) each include a single item, a 15-cm visual analog scale (VAS).

Response options/scale. The original C-HAQ DI is scored on a 4-point categorical scale that indicates how much difficulty a child has had performing each task during the past week (0 = without any difficulty, 1 = with some difficulty, 2 = with much difficulty, and 3 = unable to do). The respondent marks not applicable if the task is beyond the child's developmental age, and the respondent indicates if aids/devices or assistance were needed for any task.

Several revised versions ask respondents to compare the child's capabilities to other children the same age. The VAS$_{\text{C-HAQ38}}$ (9) uses a 10-cm VAS anchored at the left end with the phrase "Much worse than most other kids my age" and at the right end with "Much better than other kids my age." The CAT$_{\text{C-HAQ38}}$ (9) and C-HAQ38CATII (10) have 5 response options: −2 = much worse, −1 = a little worse, 0 = the same, 1 = a little better, 2 = much better compared to most kids my age. The C-HAQ38CATI (10) uses the original 0–3 response scale. The Choice$_{\text{C-HAQ38}}$ (9) includes 2 sentence stems for each question, stating 1) the activity was performed by some children "really easily," or 2) but other children weren't able to perform the activity or were only able to do so "slowly or with difficulty." The respondent chooses one sentence stem and then indicates if the chosen statement was "really true" or "sort of true" for the child.

Recall period for items. 1 week.

Endorsements. The original C-HAQ is the preferred measure of physical function in the Paediatric Rheumatology International Trials Organisation core set of outcome measures for pediatric rheumatology (11).

Examples of use. Oliveira S, Ravalli A, Pistorio A, Castell E, Malattia C, Prieur AM, et al. Proxy-reported health-related quality of life of patients with juvenile idiopathic arthritis: the Pediatric Rheumatology International Trials Organization multinational quality of life cohort study. Arthritis Rheum 2007;57:35–43.

Singh-Grewal D, Schneiderman-Walker J, Wright V, Bar-Or O, Beyene J, Selvadurai H, et al. The effects of vigorous exercise training on physical function in children with arthritis: a randomized, controlled, single-blinded trail. Arthritis Rheum 2007;57:1202–10.

Van Brussel M, Lelieveld OT, van der Net J, Engelbert RH, Helders PJ, Takken T. Aerobic and anaerobic exercise capacity in children with juvenile idiopathic arthritis. Arthritis Rheum 2007;57:891–7.

Lelieveld OT, Armbrust W, van Leeuwen MA, Duppen N, Geertzen JH, Sauer PJ, et al. Physical activity in adolescents with juvenile idiopathic arthritis. Arthritis Rheum 2008;59;1379–84.

Practical Application

How to obtain. The original C-HAQ can be found on the American College of Rheumatology (ACR) web site at http://www.rheumatology.org/practice/clinical/pediatric_assessments/chaq.pdf#search=%22CHAQ%22. Revised versions of the C-HAQ are available through the authors (9,10). There is no cost.

Method of administration. By interview or self-report of children age ≥8 years. Parent reports as proxy for a child age <8 years.

Scoring. All versions are scored by hand. The original C-HAQ and revised versions provide instructions on scoring to the respondent. No instructions for handling missing values are provided.

Score interpretation. *Part I (DI).* In the original C-HAQ, scoring range is 0–3, higher scores reflect greater disability. The highest scoring item in each category determines the score for that category. If the respondent indicates the need for assistance or the use of aids/devices to perform a task, the minimum score for that category is 2. The mean score for the 8 domains makes up the DI (range 0–3).

In most revised versions of the C-HAQ, the summary score for the DI is determined by averaging the scores on all answered items (30 or 38), eliminating the domain structure, and the use of aids/devices and assistance.

In the VAS$_{\text{C-HAQ38,}}$ each item is scored as the distance from the left end of the 10-cm line to the point where the child/parent places a slash mark to indicate the child's ability compared to other children of that age. The DI is the mean score for all 38 items. Higher scores indicate better function or less disability.

The CAT$_{\text{C-HAQ38}}$ uses a 5-point scale ranging from −2 (much worse) to 2 (much better) than other children of the same age. Higher positive scores indicate better function or less disability.

The C-HAQ38CATI uses the original C-HAQ 0–3 score range; the DI is calculated as the mean of all 38 items, with higher scores indicating greater disability.

The Choice $_{\text{C-HAQ38}}$ uses a 0–4 scale, with higher scores indicating better function or less disability.

Part II (discomfort index) and Part III (health status). Each are measured on separate 15-cm VAS. The distance from the left end of the scale to the respondent's mark is measured and multiplied by 0.2 to calculate the score (range 0–3). Both the discomfort index and health status can be rescaled to 0–100 scales. Higher scores indicate greater pain and worse overall health status.

Normative values are not available for this criterion-referenced test. Dempster et al (12) reported mean scores on the DI of the original C-HAQ that represent no disability (0), mild disability (0.24), mild-to-moderate disability (0.71), and moderate disability (1.53).

Respondent burden. The original C-HAQ takes 10 minutes for the respondent to complete, the C-HAQ$_{38}$ version takes 10–15 minutes to complete.

Adminstrative burden. The original version takes ~10 minutes when administered by interview; slightly longer for C-HAQ$_{38}$ versions.

The time to score varies by version of the C-HAQ. The original C-HAQ and C-HAQ38CATI take less than 2 minutes to score the DI; additional time is necessary to measure the VAS for pain and health status. The VAS$_{\text{C-HAQ38}}$ takes ~15–20 minutes to measure all 38 VAS for the DI plus scales for pain and health status. No special training is needed to score or interpret, although it is necessary to read the original publications describing the scoring for each version of the C-HAQ.

Translations/adaptations. The original C-HAQ has been translated and validated for use in over 40 languages and cultures including Spanish, Portuguese, Italian, Dutch, Swedish, and Norwegian. Several revised versions of the C-HAQ$_{38}$ have been translated into Dutch, Swedish, Turkish, Greek, and Danish.

Psychometric Information

Method of development. The original C-HAQ (3) consists of items from the Stanford Health Assessment Questionnaire (HAQ), adapted for use in children ages 1–19 years; several new questions were added to each functional area so there was at least 1 question that was relevant to children of all ages.

The disability section assesses functions in 8 areas or subscales that describe mostly typical ADLs, although individual item and section scores are not reported. In the C-HAQ$_{38}$ versions, additional items were derived from investigators' own experience and interviews of patients with musculoskeletal conditions. The purpose was to measure strengths as well as weaknesses by comparing the child's abilities to most other children of the same age (9,10).

Acceptability. Authors do not report the reading level, however the language for most items is simple and easy to read and understand. Missing data do not appear to be a problem and have not been addressed by the authors. The original C-HAQ is known to have a ceiling effect in children with mild disease, making it difficult to measure improvement at the better end of the functional spectrum.

Reliability. *Original C-HAQ (1).* Internal consistency (Cronbach's coefficient α) was 0.94. Test–retest stability was established for parents' responses on the DI administered once in the clinic and completed by parents at home within a mean interval of 12.8 days, with an SEM of 2.1. Paired *t*-test showed no significant difference between the 2 test scores ($P > 0.9$), and a Spearman's correlation coefficient of 0.79 ($P < 0.002$) showed strong test–retest reliability. Concordance between parent and child responses on the DI was Spearman's rho = 0.84, $P < 0.001$.

C-HAQ$_{38}$ version (9). Concordance between parent and child responses on the DI showed the intraclass correlation coefficient (ICC) values ranging from 0.41–0.68.

Validity. *Original C-HAQ (3).* Content was evaluated by a panel of experts (rheumatologists, nurses, social workers, physical and occupational therapists).

Criterion validity measures: parents' responses on the DI demonstrated strong positive correlations ($P < 0.0001$ by Kendall's tau) with accepted measures of disease activity, including the Steinbrocker Functional class (0.77), number of involved joints (0.67), physician assessment of global disease (0.67), and morning stiffness (0.54).

Discriminant validity measures: control group of 22 healthy children scored 0 on both the disability and discomfort indices, indicating the ability of the C-HAQ to discriminate between children with and without juvenile rheumatoid arthritis.

C-HAQ$_{38}$ versions (9,10,13–15). Versions that included 8 more challenging items, removal of aids/devices and assistance, and elimination of the domain structure resulted in more normally-distributed data and better ability to discriminate between patients and controls compared to versions with 30 items. Relative efficiency of the revised versions ranged from 1.01 (C-HAQ38CATI) to 2.32 (VAS$_{\text{C-HAQ38}}$).

Ability to detect change. *Juvenile arthritis.* Singh et al (3) reported parent responses on the DI correlated significantly with their global assessment of the child's functional status. The C-HAQ DI was more sensitive to change than morning stiffness or active joint count.

Dempster et al (12) did not report the ability to detect change in the original C-HAQ but provided guidelines for interpreting the clinical utility of baseline and change scores on the DI. They reported the mean score on the DI that represents no disability (0), mild disability (0.24), mild-to-moderate disability (0.71), and moderate disability (1.53). The minimum important change (MIC) scores varied by disability category. The MIC for improvement was −0 (no disability), −0.13 (mild), −0.38 (mild-to-moderate), and −0.57 (moderate). The MIC for deterioration was 0.75 (no disability and mild disability), 0.63 (mild-to-moderate disability), and 0.38 (moderate disability).

Juvenile dermatomyositis (juvenile DM). Huber et al (4) reported adequate responsiveness of the C-HAQ by comparing changes in the DI and physician global assessment (10-cm VAS) in 90 children with juvenile DM. Effect size for 18 subjects who met the criterion for improvement (3 cm on the VAS) was 1.05, with a standardized response mean (SRM) of 1.20. In 72 subjects who did not meet the criterion for improvement, effect size was 0.20 and SRM was 0.32.

Feldman et al (5) reported moderate to good responsiveness of the C-HAQ DI to clinical change in children with juvenile DM undergoing specific drug therapies. Responsiveness of revised versions for use in juvenile DM has not been reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The C-HAQ is the most frequently used measure of physical function in pediatric rheumatology; the instrument measures typical ADLs expected of children ages 1–19 years that may be difficult for some children with arthritis or other chronic musculoskeletal conditions. It is included in the core set of outcome measures for clinical trials in JIA, juvenile DM, and systemic lupus erythematosus (SLE).

The C-HAQ is simple to administer as a self-report for children age ≥8 years or by parents as proxy for their child. It takes minimal time to administer and score. Groups or conditions for which the instrument may be appropriate include children with JIA, juvenile DM, SLE, and chronic musculoskeletal pain syndromes.

Revised versions with 8 more challenging items (C-HAQ$_{38}$), modified response options, elimination of aids/devices and assistance, and a simple mean score for all 38 items appear to result in a more normal score distribution, improved ability to discriminate between patients and healthy children, and greater sensitivity to change with intervention.

Caveats and cautions. The original C-HAQ suffers from a ceiling effect in children with mild disease who have few limitations in basic and instrumental ADLs, limiting its ability to demonstrate improved physical function at the higher end of the scale. The revised versions demonstrate improved psychometric attributes and may prove over time to be a better measure of physical function in children with rheumatic disease.

Clinical usability. The original C-HAQ$_{30}$ is useful in identifying activity limitations and change in physical function over time or with intervention in children with moderate to severe disease, but less so in children with mild disease. The C-HAQ is easy to administer and score and useful in both clinical and research settings. The respondent burden is minimal.

Research usability. The original and revised versions of the C-HAQ are easy to administer and score, however the strong ceiling effect of the original C-HAQ may limit its ability to detect change in physical function in children with mild disease. The C-HAQ38CATI closely adheres to the original C-HAQ and may be the most time-efficient for research purposes. However the VAS$_{\text{C-HAQ38,}}$ in which the respondent compares the child's abilities with those of others the same age, may be more meaningful to the child and provide a better measure of the child's participation in age-appropriate activities. Dempster et al (12) provide cut points to judge a child's level of disability and to detect clinically important change at each level of disability.

JUVENILE ARTHRITIS FUNCTIONAL ASSESSMENT SCALE (JAFAS)

Description

Purpose. To measure functional ability in children with rheumatic disease, ages 7–16 years by observing the child's actual performance of typical activities under standardized conditions and procedures. It was originally published in 1989 by Lovell et al (16) as a companion measure to the Juvenile Arthritis Functional Assessment Report, a questionnaire-based measure of the child's ability to perform activities during the previous week. There have been no revisions or modifications to the original instrument.

Content. Activities of daily living (ADLs) considered by expert opinion to be important and often difficult for children with arthritis include dressing, cutting food, getting in and out of bed, picking up an object from the floor while standing, moving from standing to the floor and returning to standing, walking 50 feet unaided, and walking up a flight of 5 steps.

Number of items. 10 items in a single scale.

Response options/scale. Categorical (0–2) scale, based on the time in seconds the child takes to complete each task. The child's performance is compared to a criterion time (mean plus 2 SDs) established for healthy children (0 = task performed in less than or equal to the criterion time; 1 = task performed in time in excess of the criterion time; 2 = child unable to perform the activity).

Recall period for items. Performance-based measure (capacity).

Endorsements. Developed for use in a US Bureau of Maternal and Child Health and Resources Development project.

Examples of use. Bekkering WP, ten Cate RT, van Suijlekom-Smit LW, Mul D, van der Velde EA, van den Ende CH. The relationship between impairments in joint functions and disabilities in independent function in children with systemic juvenile idiopathic arthritis. J Rheumatol 2001;28:1099–105.

Sircar D, Ghosh B, Ghosh A, Hildar S. Juvenile idiopathic arthritis. Indian Pediatr 2006;43:429–33.

Practical Application

How to obtain. The JAFAS, standardized testing procedures, and detailed descriptions of the required equipment, performance criteria, and guidelines for timing are available from the original authors (16).

Method of administration. A trained tester observes the child's performance of each task and enters either the time (seconds) needed to complete the task or "unable to complete." Standardized testing procedures and detailed descriptions of required equipment and guidelines for testing are available from the authors (16).

Scoring. The JAFAS is scored by hand. Specific scoring instructions are found in the original paper and are available from the authors (16). The score range for each item is 0–2, and 0–20 for the entire scale (10 items). The authors do not provide instructions for handling missing values, however the scoring criteria include a score for the child being unable to complete the task.

Score interpretation. Higher scores on individual items (range 0–2) and on the full scale (range 0–20) indicate greater activity limitation. The JAFAS is a criterion-referenced measure, with the criterion (mean plus 2 SDs) being the time required by healthy children to perform each task.

Respondent burden. Approximately 10 minutes.

Adminstrative burden. Approximately 15 minutes; training to administer the test is required but is minimal.

Translations/adaptations. The original version is in English. The measure has not been adapted for other languages or cultures. Although the JAFAS was originally developed for children with juvenile rheumatoid arthritis (JRA), it is useful for children with other types of arthritis or other childhood rheumatic diseases that negatively impact a child's ability to perform basic ADLs.

Psychometric Information

Method of development. Items were selected from other assessment tools, including the Arthritis Impact Measurement Scale, Health Assessment Questionnaire, and the McMaster Health Index Questionnaire. Selected items were reviewed by a panel of pediatric occupational and physical therapists who had experience with patients with JRA; the panel chose a variety of tasks that required the use of all joints and muscle groups, were simple to test, objectively measureable, and important to daily function. The original scale of 23 items was reduced to include 10 tasks after all subjects were tested and data analysis performed (16).

Acceptability. The scale, directions, and scoring requirement are clearly written and understandable. Although not discussed by the authors, missing data do not appear to be a problem. There does appear to be a floor effect whereby children with mild disease and few limitations score very low on the scale and have little room to show improvement in their physical function with intervention.

Reliability. Cronbach's alpha for internal consistency was 0.85 (16). Interitem correlation for the JAFAS scores in children with JRA was 0.36, suggesting items measure different factors of the activity component of the International Classification of Functioning, Disability and Health. No information is available on intratester or intertester reliability or test–retest stability of the JAFAS score in children. A recent study (17) comparing the measurement properties of the JAFAS and the Childhood Health Assessment Questionnaire (C-HAQ) in a single sample of 28 children with juvenile idiopathic arthritis reported the internal consistency of the JAFAS (0.91) to be lower than the C-HAQ (0.96).

Validity. *Content validity.* Items were selected from 3 validated measures of physical function for adults with arthritis and reviewed by a panel of experts in JRA.

Convergent validity. Demonstrated by examining the correlation between JAFAS scores of children with JRA and accepted measures of disease activity; active joint count ($r = 0.40$, $P = 0.003$), Steinbrocker functional class ($r = 0.59$, $P = 0.001$), global disease activity score (0 = active, 1 = partial remission, 2 = total remission) ($r = -0.32$, $P = 0.01$). Bekkering et al (17) reported higher correlations between standard measures of disease activity and the C-HAQ ($r_s = 0.41–0.73$) then the JAFAS ($r_s = 0.07–0.50$). One reason may be the JAFAS measures the speed of a child's actual performance of tasks at a specific time point whereas the C-HAQ measures the amount of difficulty the child had performing tasks over the previous week.

Ability to detect change. There is no published evidence of the ability of the JAFAS to detect change in a child's physical function over time or with intervention.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The JAFAS is a quick and relatively simple test to administer. It addresses the activity component of the International Classification of Functioning, Disability, and Health by measuring the child's capacity to perform typical and necessary daily physical tasks. The total JAFAS score demonstrates an acceptable relationship with several measures of disease activity and severity, including swollen joint count and limited joint motion. The test may be appropriate to evaluate rehabilitation interventions in children with active disease and limited joint motion that negatively impacts their ability to perform basic ADLs. The JAFAS may be most appropriate for children whose joint range of motion is limited due to acute or chronic joint inflammation and muscle weakness.

Caveats and cautions. The JAFAS measures the child's capacity to perform only 10 simple activities and does not include complex motor skills required for individual or team sports. Thus it may not be sensitive enough to detect functional limitations that negatively impact a child's participation in school and community settings. This problem will be most evident in children with mild disease and few limitations in basic ADLs. The interrater and intrarater reliability of the JAFAS and the stability of test results over time have not been examined. Also, the ability of the test to detect change in a child's physical function over time or with intervention has not been examined.

Clinical usability. The JAFAS takes minimal training, equipment, and time to administer and score, thus it does not pose a burden on either the child or tester. It can be easily performed in a clinic or office. Clinicians can compare a child's total JAFAS score and scores on individual test items to the criterion values for healthy children in order to determine the need for a referral to physical or occupational therapy for a more detailed evaluation and intervention.

Research usability. The JAFAS demonstrates adequate validity and reliability as a measure of physical function in children with arthritis. It is also the only measure in pediatric rheumatology that measures a child's physical capacity to perform daily activities. However, the usability of the JAFAS in research is limited by the lack of information on the test's responsiveness to changes in disease status over time or to medical or rehabilitative intervention.

PEDIATRIC OUTCOMES DATA COLLECTION INSTRUMENTS (PODCI)

Description

Purpose. To assess patients under the age of 19 years for overall health, pain, and ability to participate in normal daily activities, as well as more vigorous activities typically associated with young people. The target populations are children and adolescents ages 2–18 years with general health problems, specifically any problems related to bone and muscle conditions. Originally published in 1994 (18), and updated in 2005.

Content. Five scales provide a broad view of the physical, mental, and psychosocial status of the child and adolescent patient. The scales are upper extremity and physical function, transfer and mobility tasks, sports/physical functioning, pain/comfort, treatment expectations, happiness, and satisfaction with symptoms. A Global Functioning scale consists of the mean of the "mean of items" values for the first 4 scales.

Number of items. Total measure: Adolescent Outcomes Questionnaire (self report for youth 11–18 years of age) (83 items); Adolescent Outcomes Questionnaire (parent report to be completed by a parent or guardian of a youth 11–18 years old) (86 items); and Pediatric Outcomes Questionnaire (parent report to be completed by a parent or guardian of a child ≤10 years) (86 items). Number of items in the subscales: upper extremity and physical function (8 items), transfers and mobility (11 items), sports and physical function (21 items), pain/comfort (3 items), and happiness (5 items).

Response options/scale. Most items use a categorical scale, with a range of 3–6 choices; some items require respondent to circle "yes" to all responses that apply to the patient.

Recall period for items. One week for most items; one item asks parents to indicate how often over the past 12 months the child has missed school because of health problems.

Endorsements. American Academy of Orthopedic Surgeons (AAOS)

Examples of use. Damiano D, Gilgannon M, Abel M. Responsiveness and uniqueness of the Pediatric Outcomes Data Collection Instrument to the Gross Motor Function Measure for measuring orthopedic and neurological outcomes in cerebral palsy. J Pediatr Orthop 2005;25:641–5.

Huffman GR, Bagley AM, James MA, Lerman JA, Rab G. Assessment of children with brachial plexus birth palsy using the Pediatric Outcomes Data Collection Instrument. J Pediatr Orthop 2005;25:400–4.

Lerman JA, Sullivan E, Haynes R. The Pediatric Outcomes Data Collection Instrument (PODCI) and functional assessment in patients with adolescent or juvenile idiopathic scoliosis and congenital scoliosis or kyphosis. Spine 2002;27:2052–7.

Vitale MG, Levy D, Moskowelijns A, Spellman M, Verdisco L, Roye D. Capturing quality of life in pediatric orthopeadics: two recent measures compared. J Pediatr Orthop 2001;21:629–35.

Practical Application

How to obtain. Available at no cost on the AAOS web site http://www.aaos.org/research/outcomes/outcomes_peds.asp.

Method of administration. Parent completes the Pediatric Outcomes Questionnaire as proxy for a child, ages 2–10 years, and Adolescent (parent report) Outcomes Questionnaire as proxy for adolescents, ages 11–18 years. Adolescents, ages 11–18 years, complete the Adolescent (self-report) Outcomes Questionnaire.

Scoring. Specific instructions are provided for answering each item in the questionnaires. The majority of items are scored using a 1–5 range with 1 indicating the most positive response (i.e., The activity is EASY for the child; The child is VERY HAPPY; The child NEVER required help from another person). Some items include a 6th response choice, "Child is too young." These items are coded as missing and omitted from the mean score. The AAOS web site provides a scoring worksheet in Microsoft Excel 2003 to record raw scores for all scales. General instructions are found on the Instructions tab. After administering the questionnaire, one enters the data into the worksheet on the Data Entry tab. Items with no entry are scored as missing and omitted from the mean score for that scale. Each worksheet includes formulae that build in any necessary item recoding, computation of missing items, and known general population means and SDs, as needed. Raw scores for each scale are converted to a standard score based on the mean of items that make up that scale. All items in a scale are first recalibrated so they are in the same metric, with a range of values from 0–5 for each item. Next the scores for all items comprising a scale are averaged over the number of items answered. The mean of the rescaled values is then multiplied by a constant so that each scale has a final range of values between 0–100.

Score interpretation. All standardized scores (range 0–100) are calculated in the worksheets such that higher scores represent less disability and better functioning. However, the user must exercise caution in interpreting the meaning of an individual scale score. Although the standardized scores are all in the range of 0–100, the interpretation of a single standardized score is not consistent between scales due to differences in how the general, healthy population is scored. To make the standard scores comparative across various scales, data from the general US population was transformed for each scale so that the normative score for each scale has a mean of 50 and SD of 10. Thus, a patient scoring above 50 on a particular scale is above the general population's average, while a patient scoring below 50 on a scale is below the general, healthy population's average. To compute the individual normative score requires knowledge of the general population mean (standardized) score and corresponding SDs. These values are included in the instrument's scoring worksheets and can also be found at http://www.aaos.org/research/outcomes/outcomes_documentation.asp#scoring. The normative score for an individual patient is calculated using the actual mean and SD of the 0–100 scale from the general, healthy population using the following formula: subtract the general population standardized mean from each individual's standardized score; divide this by the general population's SD; multiply the resulting value by 10 and add 50 to the resulting number. This is the final normative score for that patient. The AAOS web site provides a clear example of the formula.

Respondent burden. The time to complete is not specified on the web site but an estimate is approximately 15 minutes to complete all items. Although the exact reading level is not stated, the language is fairly simple and should be understandable for most respondents. Most items require objective responses, however a few may be emotionally sensitive to parents or adolescents; for example, "Is it easy or hard for you (your child) to make friends with children his/her own age?"

Administrative burden. Minimal time to administer; scoring must be done using the Excel worksheet on the AAOS web site and may be time-consuming. Training is not necessary, but one must follow the algorithms or formulae on the scoring worksheets to recode items to a single metric in order to calculate individual standardized scores.

Translations/adaptations. Korean (19) and Spanish (20) versions are available.

Psychometric Information

Method of development. The Pediatric Outcomes Instrument Development group (pediatric orthopedists, pediatric rheumatologists, and general adult orthopedists) agreed upon important domains to measure; these included upper and lower extremity function, ability to perform age-appropriate activities of daily living (ADLs), including recreation, pain or comfort, general happiness, expectations for treatment, and satisfaction with care. Questionnaire items were selected based on existing instruments, concerns of experts, and pilot testing with 112 parents and 64 adolescents (18). Clinical judgment, combined with classic psychometric methods (principal factor analysis, internal reliability, and item-total correction analysis) was used to create subscales that represent the domains of interest. Scales were created by averaging the item scores. Revisions to the original version were made to assess higher levels of physical function, including competitive sports, and to include response options for all items in each age group. The revised version was pilot tested in 30 parents and 30 adolescents.

Acceptability. Readability of the questionnaires is acceptable. The percent of missing items among the individual scales in the parent questionnaires (for children ages 2–10 years and adolescents ages 11–18 years) ranged from 0–25, with the most missing items in the happy and satisfied scale. Authors state that, despite missing responses on individual items, there was little impact on the creation of the scales. Most missing items were for the youngest age group (2–5 years), with the item marked as the child being too young for the activity. The authors do not report a ceiling or floor effect, however they do report an age effect in some scales, such that older patients score better than average.

Reliability. *Internal consistency.* Cronbach's alpha ranged from 0.76 (child form of happiness/satisfaction scale) to 0.95 (child form of transfers and mobility scale).

Test–retest agreement. The Pearson correlations for a subset of parents who completed a second questionnaire 1 to 2 days after the first ranged from 0.71 (happiness and satisfaction scale) to 0.97 (transfers and mobility scale).

Interrater agreement. Paired *t*-tests showed significant differences between the responses of pairs of parents and adolescents on all scales. In general, adolescents rated themselves higher on every measure of physical and mental health while parents had higher expectations for treatment outcome. The actual differences were small, considering the 0–100 score range for each scale, however the wide range of differences suggest that many parent/adolescent pairs differed significantly in their assessment of the child's condition.

Validity. Information on the validity of the PODCI can be found on the AAOS web site and in an article by Daltroy et al (18).

Content validity. Questionnaire items were selected based on their importance to patients, parents, and experts in the field; content validity was verified through pilot testing of the instruments.

Construct validity for the scales. For convergent validity, physician ratings of global function were moderately to highly correlated to parent and adolescent ratings on the global function scale (r = 0.76), upper extremity function (r = 0.62), and transfers and mobility (r = 0.75).

For divergent validity, physician ratings of patients' function and severity of diagnosis were not correlated to parent or adolescent ratings of comfort, happiness, or expectations.

For discriminant validity, 3 PODCI scales and the Child Health Questionnaire (CHQ) scales of physical function were compared, using independent *t*-tests, in 390 patients grouped by diagnosis into those characterized by upper extremity or lower extremity dysfunction, or both, omitting those whose diagnosis was not clearly associated with a specific region. Several scales discriminated between patients with and without lower extremity dysfunction, but the PODCI sports and physical function scale was the strongest. Only the PODCI upper extremity function scale discriminated between patients with and without upper extremity problems. In multiple regression analyses, 59% of the physician's assessment of global function was characterized by PODCI function scales plus the comorbidity score, while the parent's overall assessment was based more on their perception of the child's overall happiness and comorbidities.

Ability to detect change. Sensitivity of the PODCI scales to detect change over time was evaluated in several ways: 1) By correlation of change scores on the parent and physician questionnaire with parent and physician scores on a 5-point transition scale indicating change in the child's health status (much better to much worse) over a 9-month time period. Results indicated both parent and physician change score were almost completely uncorrelated with their transition scores. Higher correlations were found between change scores on the adolescent self-report and adolescent and physician transition scores. Adolescent PODCI global scores and two subscale scores correlated better with the adolescent and physician transition scores than the CHQ. However, based on regression analysis to understand the basis of parents' and adolescents' assessment of overall change in the patient's global health and orthopedic condition, the authors concluded that much of parent and adolescent judgment of improvement or decline in function was unrelated to self-perceived measures of current function or change from baseline. Parents and adolescents appear to base their judgments of change on different areas; 2) by sensitivity to change among a subset of patients who should change the most from baseline to followup assessment (score of ≤80 on a composite score of 0–100, consisting of physician-rated function, PODCI global function, and CHQ physical function). Comparing outcomes in a sample of 113 subjects with data on all key patient-report outcomes scales using *t*-tests, all PODCI scales, except upper extremity function, ranked higher on sensitivity than the CHQ physical function scale. The PODCI global score was 2.9 times more efficient than the CHQ at detecting change in this sample. In 49 adolescents with baseline composite scores of ≤80 and complete data for key scales, the most sensitive scale was the PODCI physical sports and activities scale, followed closely by the CHQ physical function and PODCI global score; and 3) By *t*-scores for sensitivity based on the severity of baseline diagnosis. Change scores were greatest for patients with diagnoses rated as most severe (n = 34). PODCI global score and 2 subscales were more sensitive to change than the CHQ physical function scale. The PODCI upper extremity function scale discriminated best between subjects with and without upper extremity impairments, 6.8 times more efficient than the CHQ global function. PODCI transfers and mobility scale was the best discriminator of function between groups with and without lower extremity impairments, being 2.6 times more efficient than the CHQ physical function scale.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PODCI scales measure important elements of activity and participation (basic and instrumental ADLs and both recreational and competitive sports) that may be negatively affected by disease. The instrument shows adequate sensitivity to detect change in physical function with interventions. The PODCI scales have been used extensively in children with cerebral palsy and other orthopedic disorders. Children with juvenile arthritis were included in the development of the PODCI, although the instruments have not been used to examine the effectiveness of interventions in this population.

Caveats and cautions. The scoring of the PODCI scales requires the use of a computer and the formulae for recoding some items can be complicated and time-consuming. Sensitivity to change appears to be strongest in patients with the most severe conditions for which there may be problems. Daltroy et al noted age effects in some scales and strongly recommended controlling for age when using the scales to compare outcomes in groups (18). Further testing specifically in groups of children with rheumatic disease should be done to confirm the instrument's usefulness in this population.

Clinical usability. Psychometric properties of the PODCI may support the interpretation of scores to make decisions for individual patients. Although the respondent burden is minimal, the requirement for using a computer program to score and interpret the scale scores and global score will limit the usefulness of the instrument in the clinic.

Research usability. The PODCI demonstrates appropriate psychometric properties to support use in clinical research. Time for the respondent to complete the questionnaire is not prohibitive (10–18 minutes). Although scoring of the individual scales and global scale requires the use of a computer, this should not limit their use in research.

ACTIVITIES SCALE FOR KIDS (ASK)

Description

Purpose. Measure physical disability and monitor functional change in children ages 5–15 years experiencing limitations in physical activities due to musculoskeletal disorders. Originally developed in 1995 and revised in 2007.

Content. The ASK capability (ASK_c) measures activities the child "could have done" (capability) and the ASK performance (ASK_p) measures activities the child actually "did do" over the past week. Items are organized in 7 domains that measure basic and instrumental activities of daily living and play in children: personal care, dressing, other skills, locomotion, play, standing, and transfers.

Number of items. Each version includes a total of 30 items: personal care (3 items), dressing (4 items), other skills (4 items), locomotion (7 items), play (2 items), standing (5 items), and transfers (5 items). Six other information items ask about the use of assistive devices and the amount of assistance the child needs for activities; these items are not included in the summary score for the scale.

Response options/scale. 5-point ordinal scale (range 0–4). The response options depend on the ASK version: ASK_c (0 = with no problem, 1 = with a little problem, 2 = with a moderate problem, 3 = with a big problem, 4 = I could not); ASK_p (0 = all of the time, 1 = most of the time, 2 = sometimes, 3 = once in a while, 4 = none of the time).

Recall period for items. Previous week.

Endorsements. The ASK is not specifically endorsed by any entity or organization, however, the instrument has been used extensively in research involving children with orthopedic disorders.

Examples of use. Moreau NG, Simpson KN, Teefy SA, Damiano DL. Muscle architecture predicts maximum strength and is related to activity levels in cerebral palsy. Phys Ther 2010;90:1619–30.

Young NL, Varni JW, Snider L, McCormick A, Sawatzky B, Scott M, et al. The internet is valid and reliable for child-report: an example using the Activities Scale for Kids (ASK) and the Pediatric Quality of Life Scale (PedsQL). J Clin Epidemiol 2009;62:314–20.

Dillon ER, Bjornson KF, Jaffe KM, Hall JG, Song K. Ambulatory activity in youth with arthrogryposis: a cohort study. J Perdiatr Orthop 2009;29:214–7.

Practical Application

How to obtain. Available for purchase from the ASK developers at: http://www.activitiesscaleforkids.com/. Contact information for the test developers is listed on the web site. The cost depends on the proposed use of the instrument (free for student projects and teaching purposes;

$115 [Canadian] for clinical site annual license or nonfunded academic research; $585 [Canadian] for funded academic research; $875 [Canadian] for multisite funded academic research).

Method of administration. Mailed to the child's home to be completed by the child in the home because most activities important to the child occur in the home, school, or community. Parents may read the items to a child younger than 9 years or those with cognitive impairments, however the child is expected to record the response.

Scoring. A test booklet for each version (ASK_p and ASK_c) includes instructions in simple, concrete language, a scale with all possible responses and the meaning of each response, and an instruction card the child can use while he/she completes each item. Children are told to complete all items. The 5-point (0–4) ordinal scale is scored as follows: 4 points to the 1st response (0), 3 points to the 2nd response (1), 2 points to the 3rd response (2), 1 point to the 4th response (3), 0 points to the 5th response (4). A 6th response option, not applicable (N/A), is not included in the computation for the summary score. Scores on the 30 individual activity items (or the number completed by the child minus any N/A items) are aggregated into a single summary score for each version by averaging the responses, then multiplying by 25 in order to convert the score to a 0–100 range, where higher scores indicate better functional outcome. Examples of the calculation are provided in the test manual.

Score interpretation. Higher scores indicate less disability or better functioning. Interpretation also depends on the purpose of the assessment. If the purpose is assessment of physical disability status, the clinician can compare a child's scores to data collected at the Hospital for Sick Children on a sample of 200 children with musculoskeletal disorders. The manual provides summary data for this sample. If the purpose is monitoring change in physical function, a change of ≥5 points is considered statistically significant. Summary scores changed by 1.73 SDs in a subsample of 34 children who were predicted by clinicians to make clinically important change. If the purpose is to identify discrepancies between capability and performance, then this may help the clinician choose areas of function on which to focus further assessment or therapy.

Normative scores are available in an article by Plint et al, where they examined the summary scores on the ASK_p for 122 healthy children (21). The mean ± SD score of 92.12 ± 6.45 (range 74.14–100) differed significantly from that of a sample of children with mild (85.86 ± 13.17), moderate (52.66 ± 22.53), and severe (21.00 ± 10.23) disabilities.

Respondent burden. The ASK takes approximately 30 minutes to complete the first time, but as little as 10 minutes on subsequent administrations. Instructions, items, and item responses are written in simple, concrete language. It may take slightly longer if a parent needs to read the items for a child younger than 9 years of age.

Adminstrative burden. No training is necessary to use the ASK, however the user must read the manual and follow directions for administration and scoring. The ASK is mailed to the child's home and is self-administered by the child. A parent may read the items to the child younger than 9 years of age, but the child must record his or her response. Approximate time to score the instrument is 15 minutes.

Translations/adaptations. The ASK web site (http://www.activitiesscaleforkids.com/private/ASK-500.html) indicates that instruments are available in several languages, including Canadian English, Canadian French, UK English, Spanish, and Dutch. Only the Canadian English version has been validated.

Psychometric Information

Method of development. The ASK was developed in 8 stages. The initial stage involved item generation by children, ages 5–15 years, with diagnoses resulting in musculoskeletal impairments (amputations, athrogryposis, arthritis, dermatomyositis, cerebral palsy, fracture, muscular dystrophy, scoliosis, spina bifida) and their parents. Children were identified through various clinics at the Hospital for Sick Children (rheumatology, orthopedic, physical therapy) and through sports camp staff. Children with significant cognitive impairment were excluded. Parents and children were interviewed in their home and asked to list activities that were difficult for them (their child) to perform and tasks the child had to stop or had never performed. Clinicians also identified items for the scale. Children ranked items in order of importance. Participants were also asked to identify and rank items from other disability scales.

Item reduction based on consensus by a panel of experts using parent and child rankings resulted in 76 items, organized into 16 sub-domains. The test was formatted with simple, concrete language in the first person and pilot tested with 10 children. Two versions were developed, the ASK_p and ASK_c and field tested with 28 children and their parents (intertester agreement). Parents and children completed the ASK 2 weeks later to assess test–retest agreement. The ASK summary scores were also compared between groups of children that referring clinicians had rated as having mild, moderate, or severe disability, showing significant differences among the groups ($P < 0.0001$). Further reduction of items was done through expert consensus based on data from an exploratory Rasch analysis (22). The original ASK versions were reduced from 73 to 53 items by removing items that did not detect disability, were not identified as important to children, or were identified as being performed less than daily or to have always met the child's needs. The 53-item versions were used to measure validity and responsiveness. Additional Rasch analysis identified the best combination of items that maximized their spread along the difficulty continuum, maximized discrimination among children across the 5–15 year age range, and provided the strongest evidence for validity and responsiveness. This resulted in 29 items plus 1 additional item (hand writing) organized into 9 domains. These were reorganized into 7 domains producing the current 30-item versions of the ASK_p and ASK_c.

Acceptability. The reading level is not stated, however the authors state the test was developed using simple, concrete language in the first person, making it easy for most children to understand. For children younger than the age of 9 years, parents may read the items to the child, although the child must record his or her response in the booklets.

The instructions state the child should complete all items. However, there is an equivalent scoring formula in the manual that takes into account the possibility of missing items (sum of all completed items times 25 divided by the maximum possible score for those items). Questions not answered or that are marked N/A are not included in either the numerator or denominator of the formula.

A ceiling effect was found in healthy children without a musculoskeletal disorder (N = 122), with most scores clustered around the higher end of the scale. No ceiling effect has been reported for either ASK version in children identified as having mild, moderate, or severe disability.

Reliability. ASK forms were administered by mail to 40 children and their parents. Upon return of the first questionnaire, a second form was mailed to the families to determine test–retest reliability. Twenty-eight children completed the first questionnaire, 18 completed the second (2).

Internal consistency. Chronbach's alpha is 0.99.

Test–retest reliability. Intraclass correlation coefficient (ICC) 0.97 for ASK_p and ICC 0.98 for ASK_c.

Interrater reliability. (Child versus parent) ICC 0.96 for ASK_p, ICC 0.98 for ASK_c.

Intrarater reliability. Child ICC 0.97 for ASK_p, child ICC 0.98 for ASK_c. Parent ICC 0.94 for ASK_p, parent ICC 0.95 for the ASK_c.

Validity. *Content validity.* The ASK was developed through generation of items from relevant literature, children with disabilities, their parents, and clinical experts. A panel of experts selected final items using a consensus approach based on the frequency of items generated by parents and children and their ranking of items for importance. Kappa coefficient (0.70) indicated items generated by parents and children were similar; Wilcoxon's signed rank test ($P = 0.055$) indicated a significant difference in the level of importance given to each activity item.

Construct validity. This was examined using several methods in a sample of 200 children (mean ± SD age 10.1 ± 3.1 years, range 5–15 years). Convergent validity was assessed. Spearman's correlation between the ASK and the Childhood Health Assessment Questionnaire (C-HAQ) was 0.82 for the ASK_p and 0.85 for the ASK_c indicating both instruments were measuring the same construct, physical disability. Divergent validity assessed through Spearman's correlation between the ASK_p and the Health Utilities Index constructs of emotion and speech were 0.15 and 0.09, respectively, which indicates these instruments are measuring different constructs. Discriminant validity was determined in a study of 200 children (mean age 10.1 years, range 5–15 years of age). Significant differences were found in the ASK_p summary scores between children at different levels of disability according to their clinicians' global ratings (A analysis of variance $P < 0.0001$) (23,24). Criterion validity was determined in a sample of 24 children (mean age 10.2 ± 3.3 years) with musculoskeletal disorders. The correlation between child and physician ASK_c scores was high, Spearman's rho = 0.92 (95% confidence interval 0.82–0.97). Interrater reliability was also high, ICC 0.99.

Ability to detect change. Responsiveness to change was assessed in 22 children chosen from the larger sample of 200; these children were expected to change during a 6-month period based on predictions by their clinicians. Parents and children completed an initial ASK_p, ASK_c, and the C-HAQ. Parents were contacted after 6 months to determine if the expected change had occurred and were sent a second set of questionnaires to complete. The standardized response mean for the 3 instruments was 1.1 (ASK_p), 0.94 (ASK_c), and 0.96 (C-HAQ), indicating the ASK_p was 16% more responsive than the C-HAQ, while the ASK_c was 2% less responsive than the C-HAQ (24).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The ASK measures a child's capability and performance of functional activities, a primary component of the International Classification of Functioning, Disability, and Health framework that may be impacted by chronic rheumatic disease. The instruments have undergone rigorous development and psychometric testing and demonstrate excellent reliability, validity, and responsiveness to change over time or with intervention. The ASK was developed to assess physical disability in children with musculoskeletal impairments due to orthopedic disorders. It is appropriate for use in children with juvenile idiopathic arthritis or other rheumatic diseases.

Caveats and cautions. Although the ASKp has been shown to discriminate between levels of disability and between children with and without musculoskeletal disability, it is not appropriate as a measure of physical disability in children without musculoskeletal impairments due to ceiling effects in this population. This may limit the usefulness of the ASK in children with rheumatic disease who do not demonstrate impairments of the musculoskeletal system. An additional limitation is that the ASK, similar to other instruments that assess physical function in children, does not elicit the child's perspective on his or her abilities relative to other children of the same age, an important aspect of participation.

Clinical usability. The ASK could be used to make decisions for individual patients regarding the focus of further assessment or therapy to address activity limitations. Because the ASK is mailed to the parent and completed in the home, there is very little administrative burden. Time for the child to compete the ASK is not excessive.

Research usability. The psychometric properties of the ASK support its use in research. The instrument has been used extensively in research involving children with primary orthopedic disorders and those with musculoskeletal impairments secondary to neurologic conditions. The ASK requires little time for respondents to complete and minimal administrative time or effort since it is mailed to respondents and completed in the home. Scoring of each ASK version is simple and requires little time.

AUTHOR CONTRIBUTIONS

Dr. Klepper drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Holsbeeke L, Ketalaar M, Schoemaker MM, Gorter JW. Capacity, capability, and performance: different constructs or three of a kind. Arch Phys Med Rehabil 2009;90:849–55.
2. Young NL, Yoshida KK, Williams JI, Bombardier C, Wright JG. The role of children in reporting their disability. Arch Phys Med Rehabil 1995; 76:913–8.
3. Singh G, Athreya BH, Fries JF, Goldsmith DP. Measurement of health status in children with juvenile rheumatoid arthritis. Arthritis Rheum 1994;37:1761–9.
4. Huber AM, Hicks JE, Lachenbruch PA, Perez MD, Zemel LS, Rennebohm RM, et al, for the Juvenile Dermatomyositis Disease Activity Collaborative Study Group. Validation of the Childhood Health Assessment Questionnaire in the juvenile idiopathic myopathies. J Rheumatol 2001;28:1106–11.
5. Feldman BM, Ayling-Campos A, Luy L, Stevens D. Silverman ED, Laxer RM. Measuring disability in juvenile dermatomyositis: validity of the Childhood Health Assessment Questionnaire. J Rheumatol 1995;22: 326–31.
6. Meiorin S, Pistorio A, Ravelli A, Iusan SM, Filocamo G, Trail L, et al. Validation of the Childhood Health Assessment Questionnaire in active juvenile systemic lupus erythematosus. Arthritis Rheum 2008;59: 1112–9.
7. Flato B, Aasland A, Vandvik IH, Forre O. Outcome and predictive factors in children with chronic idiopathic musculoskeletal pain. Clin Exp Rheumatol 1997;15:569–77.
8. Alman BA, Bhandari M, Wright JG. Function of dislocated hips in children with lower level spina bifida. J Bone Joint Surg Br 1996;78: 294–8.
9. Lam C, Young N, Marwaha J, McLimont M, Feldman BM. Revised versions of the Childhood Health Assessment Questionnaire (CHAQ) are more sensitive and suffer less from a ceiling effect. Arthritis Rheum 2004;51:881–9.
10. Groen W, Unal E, Norgaard M, Maillard S, Scott J, Berggren K, et al. Comparing different revisions of the Childhood Health Assessment Questionnaire to reduce the ceiling effect and improve score distribution: data from a multi-center European cohort study of children with JIA. Pediatric Rheumatology Online J 2010;8:16.
11. Rupeto N, Martini A. International research networks in pediatric rheumatology: the PRINTO perspective. Curr Opin Rheumatol 2004;16: 566–70.
12. Dempster H, Porepa M, Young N, Feldman BM. The clinical meaning of functional outcome scores in children with juvenile arthritis. Arthritis Rheum 2001;44:1768–74.
13. Takken T, van den Eijkof F, Hoijtink H, Helders PJ, van der Net J. Examining the psychometric characteristics of the Dutch Childhood Health Assessment Questionnaire: room for improvement? Rheumatol Int 2006;26:979–83.
14. Ouwerkerk JW, van Pelt PA, Takken T, Helders PJ, van der Net J. Evaluating score distributions in the revised Dutch version of the Childhood Health Assessment Questionnaire. Pediatr Rheumatol Online J 2008;6:14.
15. Saad-Magalhaes C, Pistorio A, Ravelli A, Filocamo G, Viola S, Brik R, et al. Does removal of aids/devices and help make a difference in the Childhood Health Assessment Questionnaire disability index? Ann Rheum Dis 2010;69:82–7.
16. Lovell DJ, Howe S, Shear E, Hartner S, McGirr G, Schulte M, et al. Development of a disability measurement tool for juvenile rheumatoid arthritis: the Juvenile Arthritis Functional Assessment Scale. Arthritis Rheum 1989;32:1390–5.
17. Bekkering WP, Cate RT, van Rossum MA, Vliet Vlieland TP. A comparison of the measurement properties of the Juvenile Arthritis Functional Assessment Scale with the Childhood Health Assessment questionnaire in daily practice. Clin Rheumatol 2007;26:1903–7.
18. Daltroy LH, Liang MH, Fossel AH, Goldberg MJ. The POSNA pediatric musculoskeletal functional health questionnaire: report on reliability, validity and sensitivity to change. Pediatric Outcomes Instrument Development Group. Pediatric Orthopaedic Society of North America. J Pediatr Orthop 1998;18:561–71.
19. Kwon DG, Chung CY, Lee KM, Lee DJ, Lee SC, Choi IH, et al. Transcultural adaptation and validation of the Korean version of the Pediatric Outcomes Data Collection Instrument (PODCI) in children and adolescents. J Pediatr Orthop 2011;31:102–6.
20. Wren TA, Sheng M, Bowen RE, Scaduto AA, Kay RM, Otsuka NY, et al. Concurrent and discriminant validity of Spanish language instruments for measuring functional health status. J Pediatr Orthop 2008;28:199–212.
21. Plinth AC, Gaboury I, Owen J, Young NL. Activities Scale for Kids: an analysis of normals. J Pediatr Orthop 2003;23:788–90.
22. Young NL, Williams JI, Yoshida KK, Wright JG. Measurement properties of the Activities Scale for Kids. J Clin Epidemiol 2000;53:125–37.
23. Young NL. Evaluation of paediatric physical disability and exploration of contributing factors [PhD thesis]. Toronto: University of Toronto; 1997.
24. Young NL, Williams JI, Yoshida KK, Bombardier C, Wright JG. The context of measuring disability: does it matter whether capability or performance is measured? J Clin Epidemiol 1996;49:1097–101.

Summary Table for Measures of Pediatric Function*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
C-HAQ	Measure health status and physical function in children 1–19 years/ original C-HAQ 30 items in 8 domains; C-HAQ38 8 more difficult items and no domain structure	Children ≥9 years of age complete the questionnaire independently; parent reports as proxy for child <9 years	30-item version ~10 minutes to complete; C-HAQ38 10–12 minutes; C-HAQ38; reading level not specified but appropriate for 9-year-old	~10–15 minutes to administer depending on version used; ~2–5 minutes to score by hand	In original C-HAQ and C-HAQ38CATI the DI range 0–3 and higher scores indicate greater disability; in $\text{VAS}_{\text{CHAQ38}}$: higher scores indicate better function	Strong for parent as respondent: Cronbach's α = 0.9 for internal consistency; SEM = 2.1 for test–retest; interrater agreement for parent and child r_s = 0.84	Excellent for content validity; moderate to high for criterion validity; C-HAQ38 versions demonstrate improved ability to discriminate patients from controls	Moderate to excellent for original C-HAQ depending on sample studied and base-line DI; ES = 1.05; SRM = 1.20	Measures physical functions that may be limited in children with rheumatic diseases; easy to administer, useful in clinical and research settings	Original C-HAQ strong ceiling effect that limits usefulness in children with mild disease and few activity limitations; C-HAQ38 versions show more normal distribution, less ceiling effect, and improved ability to detect change in children with mild disease
JAFAS	Measure capacity to perform 10 daily activities 7–16 year-old children with JRA	Trained tester observes and times the child as they perform each task using standardized protocol and equipment	~10 minutes	~15 minutes to administer and score test; training is required but minimal.	Items scored on 0–2 ordinal scale; total scale range 0–20; higher scores indicate greater activity limitation	Internal consistency α = 0.85; no data on test–retest or intertester reliability	Good for content validity; fair to moderate correlation to measures of disease activity in JRA	Unknown	Simple to administer and score; total score shows acceptable relationship to disease activity measures; may be useful in clinical assessment or research to measure effects of intervention in children with moderate disease and limited ability to perform basic ADLs	Measures only a few activities; has limited ability to detect activity limitations in children with mild disease; no evidence of ability to detect change over time or with intervention
PODCI	Daily activities, transfers, mobility, sports, pain, and treatment expectations	Parent completes as proxy for child (2–10 years); adolescents (11–18 years) complete independently	~15 minutes to compete; reading level not stated, but language is simple	Minimal time to administer; computer scored using Excel work-sheet on AAOS web site; time to score can be lengthy	Standardized scores range 0–100; higher scores indicate less disability. mean ± SD of normative score 50 ± 10	Internal consistency excellent (α = 0.76–0.95); parent test–retest (r = 0.71–0.97); poor interrater agreement for parent and adolescent responses on all scales	Content good; construct: moderate to strong for convergent validity (0.62–0.76); discriminant: validity varies by scale; able to discriminate between children with and without dysfunction in lower extremities	Sensitivity to change among subset of patients expected to change the most (baseline score ≤80) was good to excellent, in patients with diagnosis rated as severe	Measures important basic ADLs as well as sports in children and teens with orthopedic conditions; shows adequate sensitivity to change, especially in severe disability	Computer scoring necessary using AAOS software; re-scoring of some items necessary to calculate standard scores; sensitivity to change is strongest in groups with most severe disability
ASK	Measure physical disability and monitor changes in child's functional abilities	Mailed to child's home to be completed by child (ages 9–15 years); parent may read items to a child <9 years, but child must record answers	30 minutes to complete the first time; 10 minutes on repeat tests; simple, concrete language is used for instructions, items, and response choices	No training necessary, but user must read the manual and follow directions for administration and scoring; time to score ~15 minutes	Summary scores range 0–100; higher scores indicate better function or less disability	Internal consistency Cronbach's α = 0.99; test—retest ASK_p ICC 0.97, ASK_c ICC 0.98; interrater ASK_p ICC 0.96, ASK_c ICC 0.98	Excellent for content, convergent, divergent, and discriminant validity	SRM = 1.1 ASK_p, SRM = 0.94 ASK_c	Measures capability and performance components of physical functions impacted by orthopedic impairments in children; excellent psychometric attributes	May not be appropriate to measure physical function in children with no musculoskeletal limitations; does not elicit child's perspective on abilities relative to other children of similar age

* C-HAQ = Childhood Health Assessment Questionnaire; DI= disability index; VAS = visual analog scale; ES = effect size; SRM = standardized response mean; JAFAS = Juvenile Arthritis Functional Assessment Scale; JIA = juvenile idiopathic arthritis; JRA = juvenile rheumatoid arthritis; ADL = activities of daily living; PODCI = Pediatric Outcomes Data Collection Instrument; ASK = Activities Scale for Kids; AAOS = American Academy of Orthopedic Surgeons; ICC = intraclass correlation coefficient.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S112–S117
DOI 10.1002/acr.20623
© 2011, American College of Rheumatology

Measures of Disease Activity and Damage in Pediatric Systemic Lupus Erythematosus

British Isles Lupus Assessment Group (BILAG), European Consensus Lupus Activity Measurement (ECLAM), Systemic Lupus Activity Measure (SLAM), Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), Physician's Global Assessment of Disease Activity (MD Global), and Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index (SLICC/ACR DI; SDI)

BIANCA LATTANZI,[1] ALESSANDRO CONSOLARO,[1] NICOLETTA SOLARI,[1] NICOLINO RUPERTO,[2] ALBERTO MARTINI,[1] AND ANGELO RAVELLI[1]

INTRODUCTION

Although the presentation, clinical symptoms, and laboratory findings of pediatric systemic lupus erythematosus (SLE) are similar to those that are seen in adults, children and adolescents with SLE differ from adults in terms of the frequency and severity of disease activity and damage features as well as in the treatment approaches used by their attending physicians. Furthermore, assessment of pediatric patients with SLE should take into account the disease and physical/mental age–related issues that are associated with growth and development. For these reasons, it cannot be assumed a priori that the clinical measures developed for adults are suitable for children and adolescents. Therefore, outcome measures used in adults need to be subjected to critical evidence-based evaluation of their measurement properties in children and adolescents.

Because the general characteristics of disease activity and damage measures used in SLE are addressed in another article in this issue of *Arthritis Care & Research*, our review will focus on the available information specific to pediatric SLE and the critical appraisal of the value of each instrument to pediatric rheumatologists dealing with children and adolescents with SLE. As a general rule, in all measures scored items and abnormalities must be attributable to SLE.

BRITISH ISLES LUPUS ASSESSMENT GROUP (BILAG)

Psychometric Information

Validity. In a prospective observational 12-month study of 21 patients with systemic lupus erythematosus (SLE), the renal BILAG score was found to be able to differentiate between patients with nephritis (n = 10) and patients without nephritis (n = 11) (1).

Ability to detect change. Excellent responsiveness to change in disease activity was documented, by means of effect size, effect size index, standardized response mean, responsiveness statistic, and relative efficiency index, in a comparative study with the Systemic Lupus Erythematosus Activity Measure and the Systemic Lupus Erythematosus Disease Activity Index that involved 35 newly-diagnosed patients (2). In a study of 98 patients who were seen every 3 months for up to 7 visits (n = 623 total visits), the minimum clinically important difference (MCID) for clinically important improvement or worsening, based on physician's or parent's rating of the disease course between visits, was small. Using the standard error of measurement approach, the MCID value was 2 (3). The MCID is defined as "the smallest difference in a score of a disease measure of interest that patients perceive as beneficial and that would mandate, in the absence of side effects, a change in the patient management" (4).

Critical Appraisal of Overall Value to the Pediatric Rheumatology Community

Strengths. The BILAG is the most comprehensive of the SLE activity measures. It is the only SLE activity index that

[1]Bianca Lattanzi, MD, Alessandro Consolaro, MD, Nicoletta Solari, MD, Alberto Martini, MD, Angelo Ravelli, MD: Istituto di Ricovero e Cura a Carattere Scientifico G. Gaslini, and Università degli Studi di Genova, Genova, Italy; [2]Nicolino Ruperto, MD, MPH: Istituto di Ricovero e Cura a Carattere Scientifico G. Gaslini, Genova, Italy.

Address correspondence to Angelo Ravelli, MD, Pediatria II, Istituto G. Gaslini, Largo G. Gaslini 5, 16147 Genova, Italy. E-mail: angeloravelli@ospedale-gaslini.ge.it.

Submitted for publication February 2, 2011; accepted in revised form May 23, 2011.

aims specifically to show activity in individual organs/systems. It is the only transitional index, with each item that is present being recorded as new, the same, worse or improving, rather than just present or absent (5). Although initially the developers of the index did not intend to create a cumulative score, a numerical scoring scheme was developed for the 2004 version (6).

Caveats and cautions. With 86 items, it is the longest of the lupus activity tools. It remains to be established whether a different numeric conversion table should be developed for the pediatric population because of the differences in the extent of disease features between pediatric and adult SLE. It does not include immunologic tests.

Clinical usability. The BILAG is a reliable and valid instrument for measuring clinical disease activity in pediatric patients with SLE in standard clinical practice. It enables an accurate assessment of disease activity in individual organs/systems and detection of new activity or flare in one or more systems, and in which system(s). Furthermore, it helps determine when a change in therapy is needed. However, performing the BILAG is time consuming and requires training, which may limit its applicability in routine care.

Research usability. The BILAG is best suited when the assessment of the actual level or change over time in disease activity in individual organs/systems is the primary objective of the study. The excellent responsiveness to change seen in pediatric studies supports its use as a response measure in clinical trials in children and adolescents with lupus, particularly when the efficacy of a medication on single-organ involvement (e.g., nephritis, skin disease) is under scrutiny.

Advantages/disadvantages of the different versions of the BILAG. There are 2 versions of the BILAG: the original BILAG and the BILAG-2004 (5,6). An advantage of the 2004 version is that its numerical scoring system may overcome the inability to give an overall score in the original BILAG (7). All studies performed in pediatric SLE have used the original version of the index.

EUROPEAN CONSENSUS LUPUS ACTIVITY MEASUREMENT (ECLAM)

Psychometric Information

Validity. The ECLAM (8) was found to have construct validity and to perform similarly to the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) in predicting disease damage and the need for steroids in 66 newly diagnosed patients with pediatric systemic lupus erythematosus (SLE) (9). In a multinational study that included 557 patients who underwent a baseline visit, at the time of an active phase of disease requiring a major therapeutic intervention, and a subsequent visit after 6 months, the ECLAM was found to possess significant ability to discriminate patients who were improved or not improved at 6 months, based on the physician's or parent's assessment of the child's response to therapy. In the same study, the ECLAM was found to have good construct validity, i.e., it was moderately correlated with the Physician's Global Assessment of Disease Activity and predicted strongly the child's response to therapy (10). The ECLAM was subsequently found to predict improvement according to the evaluation of the participants in the consensus conference that led to the development of the provisional criteria for the evaluation of response to therapy in pediatric SLE (11).

Ability to detect change. The ECLAM was found to be very responsive to change in disease activity and slightly more responsive than the SLEDAI in 66 newly diagnosed patients with pediatric SLE. Responsiveness statistics included effect size, effect size index, standardized response mean (SRM), and relative efficiency index (9). In a study of 98 patients who underwent a total of 623 visits, the minimum clinically important difference (MCID) for clinically important improvement or worsening, based on physician's or parent's rating of the disease course between visits, was small. Using the standard error of measurement approach, the MCID value was 1 (3). In a multinational study that included 557 patients who underwent a baseline visit, at the time of an active phase of disease requiring a major therapeutic intervention, and a subsequent visit after 6 months, the ECLAM was found to be strongly responsive to change in disease activity (SRM 1.3). In the same study, the ECLAM was found to be slightly more responsive and less skewed than the SLEDAI (10).

Critical Appraisal of Overall Value to the Pediatric Rheumatology Community

Strengths. It has been suggested that the ECLAM may be preferable to the SLEDAI for measuring disease activity because of its superior quantitative properties (9). The ECLAM has the potential advantage over the Systemic Lupus Activity Measure and SLEDAI of using the entire range of possible scores, which means that scores are less skewed.

Caveats and cautions. It has been argued that face validity of the ECLAM may be inferior to that of the SLEDAI because items are not exactly defined, the time frame during which symptoms are regarded as "evolving" is unclear, the decrease in the complement levels necessary to be scored as "significantly reduced compared to the last observation" (attribute 12b) is not well defined, and there is a lack of a clearly defined time frame during which symptoms have to occur to be included in a certain measurement of disease activity (9).

Clinical usability. The ECLAM is reasonably short and simple, which makes it feasible for use in standard clinical practice. Factors potentially hampering its application include unclear definition of items and time frames, and complexity of calculation of the total disease activity score, which is not equal to the simple sum of the domains scores.

Research usability. The ECLAM has demonstrated good construct, discriminative and predictive ability, and excellent responsiveness to change over time in patients with pediatric SLE and is, therefore, a valid instrument for the assessment of disease activity in both clinical research and therapeutic trials.

SYSTEMIC LUPUS ACTIVITY MEASURE (SLAM)

Psychometric Information

Validity. SLAM (12) use in childhood-onset systemic lupus erythematosus (SLE) was assessed in a comparative study using the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), British Isles Lupus Assessment Group (BILAG), and SLAM in a pediatric population of 35 patients (2). In a multinational study that included 557 patients who underwent a baseline visit, at the time of an active phase of disease requiring a major therapeutic intervention, and a subsequent visit after 6 months, the SLAM was found to possess significant ability to discriminate patients who were improved or not improved at 6 months based on the physician's or parent's assessment of the child's response to therapy (10).

Ability to detect change. Excellent responsiveness to change in disease activity has been documented, by means of effect size, effect size index, standardized response mean (SRM), responsiveness statistic, and relative efficiency index in a comparative study with BILAG and SLEDAI that involved 35 newly-diagnosed patients (2). In a study of 98 patients who underwent a total of 623 visits, the minimum clinically important difference (MCID) for clinically important improvement or worsening, based on physician's or parent's rating of the disease course between visits, was small. Using the standard error of measurement approach, the MCID value was 4 (3). In a multinational study that included 557 patients who underwent a baseline visit, at the time of an active phase of disease requiring a major therapeutic intervention, and a subsequent visit after 6 months, the SLAM was found to be strongly responsive to change in disease activity (SRM 1.3) (10).

Critical Appraisal of Overall Value to the Pediatric Rheumatology Community

Strengths. The SLAM and the latest version, SLAM-R, include more systemic features than the SLEDAI. They assess subjective symptoms (fatigue, arthralgias, myalgias), which may increase their correlation with parent/patient self assessment of function and general health. All items are weighted, which enables an accurate grading of clinical and laboratory abnormalities by severity. A comprehensive set of laboratory tests is incorporated.

Caveats and cautions. Inclusion of many subjective items, which may not be related directly to disease activity, may detract from the validity of the index. The SLAM gives equal weighting to different levels of severity of organ disease activity without considering the significance of the organ involved. It does not include immunologic tests.

Clinical usability. The SLAM was found to be very user friendly in a comparative study with BILAG and SLEDAI (2). Item grading makes this index potentially more flexible than the SLEDAI for monitoring of changes in disease over time in standard clinical practice. However, the SLAM is longer and somewhat more complex than the SLEDAI.

Research usability. The SLAM has shown excellent psychometric properties in validation analyses in patients with pediatric SLE and is therefore well suited for use in clinical research and therapeutic trials.

SYSTEMIC LUPUS ERYTHEMATOSUS DISEASE ACTIVITY INDEX (SLEDAI)

Psychometric Information

Validity. The SLEDAI (13) was found to have construct validity and to perform similarly to the European Consensus Lupus Activity Measurement (ECLAM) in predicting disease damage and the need for steroids in 66 newly diagnosed patients with pediatric systemic lupus erythematosus (SLE). In a multinational study that included 557 patients who underwent a baseline visit, at the time of an active phase of disease requiring a major therapeutic intervention, and a subsequent visit after 6 months, the SLEDAI was found to possess significant ability to discriminate patients who were improved or not improved at 6 months based on the physician's or parent's assessment of the child's response to therapy (10).

Ability to detect change. Excellent responsiveness to change in disease activity has been documented, by means of effect size, effect size index, standardized response mean (SRM), responsiveness statistic, and relative efficiency index, in a comparative study with the British Isles Lupus Assessment Group (BILAG) and Systemic Lupus Activity Measure (SLAM) that involved 35 newly-diagnosed patients (2). In a study of 98 patients who underwent a total of 623 visits, the minimum clinically important difference (MCID) for clinically important improvement or worsening, based on physician's or parent's rating of the disease course between visits, was small. Using the standardized error of measurement approach, the MCID value was 2 (3). In a multinational study that included 557 patients who underwent a baseline visit, at the time of an active phase of disease requiring a major therapeutic intervention, and a subsequent visit after 6 months, the ECLAM was found to be strongly responsive to change in disease activity (SRM 1.1) (10).

Critical Appraisal of Overall Value to the Pediatric Rheumatology Community

Strengths. The SLEDAI includes only 24 items, which makes it the shortest of the lupus activity tools. The SLEDAI gives more points to renal disease than does the SLAM (up to 16 points versus a maximum of 8 points), which makes it potentially more responsive in patients who relapse with renal disease primarily.

Caveats and cautions. The SLEDAI is the only lupus activity tool that does not include subjective items, such as fatigue, joint pain, etc. This makes it potentially less suitable in capturing patient-relevant disease changes. Individual items are not graded for severity. It has been argued that the SLEDAI may not capture sufficiently worsening of an already existing feature or detect partial improvement. However, the 2000 modification of the index (SLEDAI-2K) enables recording ongoing disease activity, as well as new or deteriorating disease activity (14).

Clinical usability. The SLEDAI was found to be the quickest measure to complete in a comparative study with BILAG and SLAM (2). In the clinical setting, it may be the preferable lupus activity tool because it is concise and easy to complete.

Research usability. The SLEDAI has shown excellent psychometric properties in validation analyses in patients with pediatric SLE and is, therefore well suited for use in clinical research and therapeutic trials.

Advantages/disadvantages of the different versions of the SLEDAI. There are 3 versions of the SLEDAI: the original SLEDAI (13), the SLEDAI-2K (14), and the MEX-SLEDAI (15). In the original version, the items rash, alopecia, mucous membrane lesions, and proteinuria are scored only if they represent their first occurrence or a recurrence (or a recent increase for proteinuria), whereas in the SLEDAI-2K version these items are simply scored when present. This change in the 2K version was made to reflect ongoing disease activity in the affected organ systems. The MEX-SLEDAI has the advantage of avoiding the cost of immunologic laboratory tests because it does not include anti–double-stranded DNA antibodies and complement levels. All studies performed in pediatric SLE have used the original version of the SLEDAI, except for the study by Brunner et al (3), which was based on the SLEDAI-2K version.

PHYSICIAN'S GLOBAL ASSESSMENT OF DISEASE ACTIVITY (MD GLOBAL)

Psychometric Information

Validity. In a multinational study that included 557 patients who underwent a baseline visit, at the time of an active phase of disease requiring a major therapeutic intervention, and a subsequent visit after 6 months, the MD Global demonstrated a strong ability in discriminating patients who were improved or not improved at 6 months based on the physician's or parent's assessment of the child's response to therapy. The discriminative ability of the MD Global as well as its ability to predict response to therapy were comparable to that of the European Consensus Lupus Activity Measurement (ECLAM). The baseline-to-6 month change in the MD Global was moderately correlated with the change in the ECLAM, physical summary score of the Child Health Questionnaire, and parent global assessment of the patient's overall well-being, and poorly correlated with the change in the 24-hour proteinuria (10). The MD Global was subsequently found to predict improvement according to the evaluation of the participants in the consensus conference that led to the development of the provisional criteria for the evaluation of response to therapy in pediatric systemic lupus erythematosus (SLE) (11).

Ability to detect change. In a multinational study that included 557 patients who underwent a baseline visit, at the time of an active phase of disease requiring a major therapeutic intervention, and a subsequent visit after 6 months, the MD Global was found to be the most responsive measure, together with the ECLAM and the Systemic Lupus Activity Measure (SLAM; standardized response mean 1.3 for all 3 measures) (10). In a study of 98 patients who underwent a total of 623 visits, the minimum clinically important difference (MCID) for clinically important improvement or worsening, based on physician's or parent's rating of the disease course between visits, was small. Using the standard error of measurement approach, the MCID value was 1. The change-corrected agreement of activity index with a stable course was greater for the MD Global than for the ECLAM, SLAM, and Systemic Lupus Erythematosus Disease Activity Index (3).

Critical Appraisal of Overall Value to the Pediatric Rheumatology Community

Strengths. The MD Global is the simplest and most feasible of the physician-reported disease activity measures. Furthermore, its statistical performances were found to be comparable to those of the composite disease activity tools.

Caveats and cautions. The MD Global is a broad measure of SLE activity and, therefore, may not detect with sufficient reliability improvement or worsening of disease activity in individual organs/systems. Use of the 10-cm horizontal line visual analog scale (VAS) to rate the MD Global may lead to inaccuracies in assessing disease remission. Due to the relative aversion to extremes that is often seen when using such VAS, very low values (0.1 or 0.2 cm) are frequently obtained when the assessor actually intended to mark the end of the line. It has been suggested that the 21-circle VAS format may be less skewed and less affected by ceiling effect and has the potential advantage of increasing the accuracy of assessment of clinical remission (16).

Clinical usability. The simplicity and ease of the MD Global makes it well suited for use in monitoring the course of disease activity over time in standard clinical practice.

Research usability. The MD Global has shown excellent psychometric properties in patients with pediatric SLE and is therefore well suited for use in clinical research and therapeutic trials.

SYSTEMIC LUPUS INTERNATIONAL COLLABORATING CLINICS/AMERICAN COLLEGE OF RHEUMATOLOGY DAMAGE INDEX (SLICC/ACR DI; SDI)

Psychometric Information

Validity. Several studies have shown that the SDI is a valid and reliable instrument to capture damage in patients with pediatric systemic lupus erythematosus (SLE) (17–21). Accumulated damage, measured with the SDI, was found to be predicted by cumulative disease activity over time and to be correlated with the frequency of severe disease flares in the first 3 years of followup. Item weightings for the SDI using Rasch analysis were not found to lead to an important clinical improvement (22). Rasch analysis is a psychometric approach that has been used to assess and improve rheumatology scales by generating appropriate item weightings (23). The presence of accumulated damage, measured with the SDI, was found to affect significantly the health-related quality of life, particularly in the physical domain in patients with pediatric SLE (24).

Critical Appraisal of Overall Value to the Pediatric Rheumatology Community

Strengths. The SDI (25,26) enables a detailed and comprehensive assessment and cumulative organ/system damage in pediatric patients with SLE. It constitutes an important tool to monitor over time the development of damage due to active inflammation, medication side effects, and comorbid conditions. Regular use of the SDI ensures harmonization of long-term studies of pediatric patients into adulthood.

Caveats and cautions. It has been argued that the SDI does not cover all forms of damage that children or adolescents with SLE may develop over time, particularly effects on growth and development (27). Incorporation of growth retardation and pubertal delay in a modified pediatric version of the SDI has been advised. Furthermore, a redefinition of the item cognitive impairment has been proposed to facilitate its applicability in younger patients. Since some SDI items, such as myocardial infarction, pancreatic insufficiency, claudication, gastrointestinal stricture, ruptured tendons, and malignancy, are rarely seen in children and adolescents with SLE, the utility of their assessment in pediatric SLE has been questioned (27). To avoid confusion between active inflammation and irreversible damage, in order to be scored in the SDI an item needs to be present for 6 months (except damage items that are theoretically nonreversible). Thus, the SDI covers, by definition, only irreversible damage and does not take into account the ability of children to recover and regenerate to a greater degree than adults. For instance, avascular necrosis may have a potential for regeneration and remodeling of bone lesions in children if better control of disease without the use of steroids is achieved and if normal growth velocity is restored. Furthermore, children have an exceptional capacity for neurologic recovery that adults lack, and full recovery may occur months to years later (28). It has also been argued that since steroid-induced diabetes mellitus may be reversible, the presence of steroid-induced diabetes mellitus for 6 months or more may not represent a true permanent damage (29).

To overcome the limitations of the SDI, a modified pediatric version, (Ped-SDI) has been proposed (27). As compared with the original SDI, the pediatric version includes 2 additional items: growth failure and delayed puberty. The glossary of terms for the items of the original SDI was maintained, with the sole exception of the indication that in younger children proteinuria should be adjusted for height and weight. A specific definition for the new items growth failure and delayed puberty was provided. At variance from the original SDI, which defines damage as an irreversible change in an organ or system, it was taken into account that some forms of damage are potentially reversible in pediatric patients.

Modifications to the pediatric version of the SDI suggested subsequently are to rename the item growth failure as "reduced final height," to reflect the really irreversible outcome, and to alter the definition of pubertal delay to mean a significant lack of pubertal progression to indicate permanent damage to the hypothalamic-pituitary axis. Furthermore, the modification of the definition of the item of gonadal failure has been advised, as gonadal failure defined as secondary amenorrhea before the age of 40 years is not easily applied to adolescent girls, whose menstrual cycles may be irregular as a normal physiologic variant for the first 2 years after menarche (29).

Clinical usability. The pediatric version of the SDI enables an accurate assessment and monitoring of the main forms of cumulative organ/system damage that can occur in pediatric patients with SLE. It is simple and easy to complete and score. Regular (i.e., yearly) completion of the SDI provides clinicians with an important tool to follow the course of organ/system damage from the pediatric age into adulthood (30).

Research usability. Application of the SDI and its pediatric version in pediatric patients with SLE has shown that the index is valid for use in observational cohort studies and long-term outcome surveys. It may also be valuable in the prediction of outcome.

DISCUSSION

All global measures of disease activity have been found to be reliable and valid for use in children and adolescents with SLE, and none of them has shown clearly superior metrologic properties. The choice of a specific tool may largely depend on the purposes of the study, the investigational setting (standard clinical practice or research), or the personal preference of the investigator. Owing to its simplicity, feasibility, and good psychometric properties, the MD Global should always be incorporated in the assessment of disease activity either in standard clinical practice and research. Although the SDI has proved suitable to assess damage in patients with pediatric SLE, it was found to have some important limitations for use in the pediatric age group, the chief of which is the inability to capture some forms of damage that are unique to children and adolescents, namely growth failure and delayed puberty. Use of the modified pediatric version of the SDI in pediatric patients with SLE is, therefore, advised.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Marks SD, Pilkington C, Woo P, Dillon MJ. The use of the British Isles Lupus Assessment Group (BILAG) index as a valid tool in assessing disease activity in childhood-onset systemic lupus erythematosus. Rheumatology (Oxford) 2004;43:1186–9.
2. Brunner HI, Feldman BM, Bombardier C, Silverman ED. Sensitivity of

the Systemic Lupus Erythematosus Disease Activity Index, British Isles Lupus Assessment Group Index, and Systemic Lupus Activity Measure in the evaluation of clinical change in childhood-onset systemic lupus erythematosus. Arthritis Rheum 1999;42:1354–60.

3. Brunner HI, Higgins GC, Klein-Gitelman MS, Lapidus SK, Olson JC, Onel K, et al. Minimal clinically important differences of disease activity indices in childhood-onset systemic lupus erythematosus. Arthritis Care Res (Hoboken) 2010;62:950–9.
4. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15.
5. Symmons DP, Coppock JS, Bacon PA, Bresnihan B, Isenberg DA, Maddison P, et al. Development and assessment of a computerized index of clinical disease activity in systemic lupus erythematosus: members of the British Isles Lupus Assessment Group (BILAG). Q J Med 1988;69: 927–37.
6. Yee CS, Cresswell L, Farewell V, Rahman A, Teh LS, Griffiths B, et al. Numerical scoring for the BILAG-2004 index. Rheumatology (Oxford) 2010;49:1665–9.
7. Pope J. The revised BILAG Index with numerical scoring in systemic lupus erythematosus: added value with some limitations. Rheumatology (Oxford) 2010;49:1616–7.
8. Bencivelli W, Vitali C, Isenberg DA, Smolen JS, Snaith ML, Sciuto M, et al. Disease activity in systemic lupus erythematosus: report of the Consensus Study Group of the European Workshop for Rheumatology Research III: development of a computerised clinical chart and its application to the comparison of different indices of disease activity. The European Consensus Study Group for Disease Activity in SLE. Clin Exp Rheumatol 1992;10:549–54.
9. Brunner HI, Silverman ED, Bombardier C, Feldman BM. European consensus lupus activity measurement is sensitive to change in disease activity in childhood-onset systemic lupus erythematosus. Arthritis Rheum 2003;49:335–41.
10. Ruperto N, Ravelli A, Cuttica R, Espada G, Ozen S, Porras O, et al, for the Pediatric Rheumatology International Trials Organization (PRINTO) and the Pediatric Rheumatology Collaborative Study Group (PRCSG). The Pediatric Rheumatology International Trials Organization criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: prospective validation of the disease activity core set. Arthritis Rheum 2005;52:2854–64.
11. Ruperto N, Ravelli A, Oliveira S, Alessio M, Mihaylova D, Pasic S, et al. for the Pediatric Rheumatology International Trials Organization (PRINTO) and the Pediatric Rheumatology Collaborative Study Group (PRCSG). The Pediatric Rheumatology International Trials Organization/American College of Rheumatology provisional criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: prospective validation of the definition of improvement. Arthritis Rheum 2006;55:355–63.
12. Liang MH, Socher SA, Larson MG, Schur PH. Reliability and validity of six systems for the clinical assessment of disease activity in systemic lupus erythematosus. Arthritis Rheum 1989;32:1107–18.
13. Schned ES, Glickstein SL, Doyle MA. Derivation of the SLEDAI [letter]. Arthritis Rheum 1993;36:877.
14. Gladman DD, Ibanez D, Urowitz MB. Systemic lupus erythematosus disease activity index 2000. J Rheumatol 2002;29:288–91.
15. Guzman J, Cardiel MH, Arce-Salinas A, Sanchez-Guerrero J, Alarcon-Segovia D. Measurement of disease activity in systemic lupus erythematosus: prospective validation of 3 clinical indices. J Rheumatol 1992;19:1551–8.
16. Filocamo G, Davi S, Pistorio A, Bertamino M, Ruperto N, Lattanzi B, et al. Evaluation of 21-numbered circle and 10-centimeter horizontal line visual analog scales for physician and parent subjective ratings in juvenile idiopathic arthritis. J Rheumatol 2010;37:1534–41.
17. Brunner HI, Silverman ED, To T, Bombardier C, Feldman BM. Risk factors for damage in childhood-onset systemic lupus erythematosus: cumulative disease activity and medication use predict disease damage. Arthritis Rheum 2002;46:436–44.
18. Ravelli A, Duarte-Salazar C, Buratti S, Reiff A, Bernstein B, Maldonado-Velazquez MR, et al. Assessment of damage in juvenile-onset systemic lupus erythematosus: a multicenter cohort study. Arthritis Rheum 2003;49:501–7.
19. Miettunen PM, Ortiz-Alvarez O, Petty RE, Cimaz R, Malleson PN, Cabral DA, et al. Gender and ethnic origin have no effect on longterm outcome of childhood-onset systemic lupus erythematosus. J Rheumatol 2004;31:1650–4.
20. Lilleby V, Flato B, Forre O. Disease duration, hypertension and medication requirements are associated with organ damage in childhood-onset systemic lupus erythematosus. Clin Exp Rheumatol 2005;23: 261–9.
21. Bandeira M, Buratti S, Bartoli M, Gasparini C, Breda L, Pistorio A, et al. Relationship between damage accrual, disease flares and cumulative drug therapies in juvenile-onset systemic lupus erythematosus. Lupus 2006;15:515–20.
22. Brunner HI, Feldman BM, Urowitz MB, Gladman DD. Item weightings for the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Disease Damage Index using Rasch analysis do not lead to an important improvement. J Rheumatol 2003;30:292–7.
23. Rasch G. An item analysis which takes individual differences into account. Br J Math Stat Psychol 1966;19:49–57.
24. Ruperto N, Buratti S, Duarte-Salazar, Pistorio A, Reiff A, Bernstein B, et al. Health-related quality of life in juvenile-onset systemic lupus erythematosus and its relationship to disease activity and damage. Arthritis Rheum 2004;51:458–64.
25. Gladman D, Ginzler E, Goldsmith C, Fortin P, Liang M, Urowitz M, et al. The development and initial validation of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology damage index for systemic lupus erythematosus. Arthritis Rheum 1996;39:363–9.
26. Gladman DD, Urowitz MB, Goldsmith CH, Fortin P, Ginzler E, Gordon C, et al. The reliability of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology damage index in patients with systemic lupus erythematosus. Arthritis Rheum 1997;40: 809–13.
27. Gutierrez-Suarez R, Ruperto N, Gastaldi R, Pistorio A, Felici E, Burgos-Vargas R, et al. A proposal for a pediatric version of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index based on the analysis of 1,015 patients with juvenile-onset systemic lupus erythematosus. Arthritis Rheum 2006; 54:2989–96.
28. Dayal NA, Gordon C, Tucker L, Isenberg DA. The SLICC damage index: past, present and future. Lupus 2002;11:261–5.
29. Hiraki LT, Hamilton J, Silverman ED. Measuring permanent damage in pediatric systemic lupus erythematosus. Lupus 2007;16:657–62.
30. Ravelli A, Ruperto N, Martini A. Outcome in juvenile onset systemic lupus erythematosus. Curr Opin Rheumatol 2005;17:568–73.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S253–S262
DOI 10.1002/acr.20634
© 2011, American College of Rheumatology

Measures of Pediatric Pain

21–Numbered Circle Visual Analog Scale (VAS), E-Ouch Electronic Pain Diary, Oucher, Pain Behavior Observation Method, Pediatric Pain Assessment Tool (PPAT), and Pediatric Pain Questionnaire (PPQ)

CATRINA C. LOOTENS[1] AND MICHAEL A. RAPOFF[2]

INTRODUCTION

Chronic pain is a primary feature of juvenile arthritis (JA) (1). Patients with JA often report mild to moderate pain (2–6). Approximately 25–30% report moderate to severe pain (7,8), and most children with JA report at least some pain lasting from 30 minutes to 24 hours a day, with a mean of 4.3 hours per day (9). A 2-month daily diary study showed that children with JA report pain on an average of 73% of the days, with the majority (76%) reporting pain on >60% of the days (10). Approximately 60% of children with JA report joint pain at disease onset, 50% report pain at 1-year followup, and 40% continue to report pain 5 years later (11). Moreover, adults who as children were diagnosed with JA report significantly more pain, fatigue, and disability than sex-matched healthy controls (12). Therefore, pain is a significant problem for some children with JA that persists into adulthood and is associated with greater disability. However, the above pain references are from 1997 or earlier, which precedes the use of biologic agents. More recent studies have shown lower levels of pain (13).

In order to effectively document and treat JA-associated pain, we need reliable, valid, and clinically useful measures of pain. Both self-report and observational measures of pain have been used to measure pain in children and adolescents (14,15). However, observational measures are limited for recurrent or chronic pain, as with JA, because overt pain behaviors tend to habituate or dissipate over time (15). Therefore, self-report measures are preferable when measuring JA-associated pain, except with very young children or children with cognitive deficits. Autonomic measures (e.g., pulse) have been used in JA (16), but are not always considered proxy measures of pain (17). The following is a review of self-report and observational pain measures that have been tested with children and adolescents with JA. For measures that have been used for other pediatric acute and chronic recurring pain conditions, please see the Pediatric Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials recommendations (18). All but one of the measures included in the current review are self-report measures.

21–NUMBERED CIRCLE VISUAL ANALOG SCALE (VAS)

Description

Purpose. The 21–numbered circle VAS measures overall well-being, pain intensity, and overall disease activity in children with juvenile arthritis (JA). The 21–numbered circle VAS was originally examined in adults with rheumatoid arthritis (RA). It was recently evaluated with physician and parent ratings of children with JA.

Content. Physician and parent report using the 21–numbered circle VAS may be completed as part of a battery of assessments used to monitor symptoms of children with JA being seen in a rheumatology clinic. The domains assessed by the 21–numbered circle VAS include overall well-being, pain intensity, and overall disease activity. Each domain is assessed using a VAS with anchors at each end and 21 circles in 0.5-unit increments. When selecting a rating, the parent or physician fills in 1 of the 21 circles on the VAS.

Number of items. The 21–numbered circle VAS consists of 3 items: 1) parent global rating of child's well-being, 2) parent rating of child's pain intensity, and 3) physician global assessment of overall disease activity.

Response options/scale. Parent overall rating of child's well-being is assessed with a 21–numbered circle VAS

[1]Catrina C. Lootens, MA: University of Kansas, Lawrence; [2]Michael A. Rapoff, PhD: University of Kansas Medical Center, Kansas City.

Dr. Rapoff has received consultant fees, speaking fees, and/or honoraria (less than $10,000 each) from American College of Rheumatology, NIH, and Society of Pediatric Psychology, and receives royalties from Springer for the book *Adherence to Pediatric Medical Regimens: Second Edition.*

Address correspondence to Michael A. Rapoff, PhD, University of Kansas Medical Center, Department of Pediatrics, 3901 Rainbow Boulevard, Kansas City, KS 66103. Email: mrapoff@kumc.edu.

Submitted for publication February 22, 2011; accepted in revised form May 10, 2011.

anchored at "very well" to "very poorly," with a drawing of a face at each end. Parent rating of child's pain intensity is assessed with a 21–numbered circle VAS anchored at "no pain" to "very severe pain," with a drawing of a face at each end. Physician global assessment of overall disease activity is assessed with a 21–numbered circle VAS anchored at "no activity" and "maximum activity." There are no drawings present on this item.

Recall period for items. Recall period for the 3 domains are as follows: 1) "at this time" for parent overall rating of child's well-being, 2) "in the past week" for parent rating of child's pain intensity, and 3) "at the time of the present visit" for physician global assessment of overall disease activity.

Endorsements. None.

Examples of use. Pincus T, Bergman M, Sokka T, Roth J, Swearingen C, Yazici Y. Visual analog scales in formats other than a 10 centimeter horizontal line to assess pain and other clinical data. J Rheumatol 2008;35:1550–8 (13).

Filocamo G, Davi S, Pistorio A, Bertamino M, Ruperto N, Lattanzi B, et al. Evaluation of 21–numbered circle and 10-centimeter horizontal line visual analog scales for physician and parent subjective ratings in juvenile idiopathic arthritis. J Rheumatol 2010;37:1534–41 (19).

Practical Application

How to obtain. The 21–numbered circle VAS for pediatric rheumatology was developed by researchers in Genoa, Italy (Dr. A. Ravelli, Pediatria II, Istituto G. Gaslini, Largo G. Gaslini 5, 16147 Genoa, Italy. E-mail: angeloravelli@ospedale-gaslini.ge.it).

Method of administration. Parent and physician report.

Scoring. The values for each item on the 21–numbered circle VAS may be seen underneath each of the 21 circles. Computer scoring is not necessary.

Score interpretation. The score range for each item is 0–10. Higher values indicate poorer overall well-being, more severe pain, and more disease activity.

Respondent burden. Each of the 3 items on the 21–numbered circle VAS requires ~5–10 seconds to complete.

Administrative burden. The 21–numbered circle VAS requires 7.4 seconds to hand score.

Translations/adaptations. The 21–numbered circle VAS was originally evaluated with a sample of adults with RA. It was subsequently evaluated with a sample of children with JA receiving treatment in a rheumatology clinic in Italy.

Psychometric Information

Method of development. The American College of Rheumatology (ACR) pediatric response criteria for JA (20) were considered during the development of the 21–numbered circle VAS. Two items on the 21–numbered circle VAS (i.e., parent global rating of child's well-being and physician global assessment of overall disease activity) are included by the ACR as criteria demonstrating responsiveness in JA.

Acceptability. The authors of the only study assessing the 21–numbered circle VAS in a pediatric rheumatology population suggest respondents better understand the measure than a traditional 10-cm horizontal-line VAS. Ceiling effects were found for 32.9% and 43.7% for parent global rating of overall well-being and physician global assessment of overall disease activity, respectively (19).

Reliability. Not reported.

Validity. Construct validity for the 21–numbered circle VAS has been established. Spearman's correlations between scores on the 21–numbered circle VAS and other JA outcomes (i.e., the Juvenile Arthritis Functionality Scale, tender joint count, restricted joint count, and active joint count) ranged from 0.42–0.88. Correlations between scores on the 21–numbered circle VAS and laboratory variables (i.e., erythrocyte sedimentation rate and C-reactive protein level) were lower, with a range of 0.33–0.54. Correlations between scores on the 21–numbered circle VAS and the Pediatric Rheumatology Quality of Life scale ranged from 0.30–0.75 (19).

Ability to detect change. One study established the responsiveness of the 21–numbered circle VAS. Children in this study attended a second clinic visit at a mean ± SD of 6 ± 3 months after an initial visit. At this second visit, physicians and parents rated the child's disease course from the previous visit as: 1) "much improved," 2) "slightly improved," 3) "stable," 4) "slightly worsened," or 5) "much worsened." The standardized response mean (SRM) was calculated as the mean score change divided by the SD of the individual's score change. SRM values in children with improved disease were ~0.8, and in children with worsened disease values ranged from 0.6–0.8. SRM values in children with stable disease were ~0. Separate physician and parent minimum clinically significant difference (MCID) values were computed by calculating the mean change in score between visits in patients rated as "slightly improved" or "slightly worsened." MCID values for improvement ranged from −2.2 to −0.6, and ranged from 1.4–2.3 for worsening. MCID values for children classified as "stable" were ~0 (19).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The 21–numbered circle VAS is easy to administer and score. It has been found to perform similarly to a traditional 10-cm horizontal-line VAS. Initial evaluations suggest it may be appropriate for evaluating interventions.

Caveats and cautions. Ceiling effects of physician rating of disease activity may be a problem in children with well-controlled disease. This problem may not be unique to this measure, but common to JA due to use of increasingly effective treatments (e.g., use of biologic medications). Additional psychometric evaluations, particularly reliability assessments, are needed. Current studies do not include evaluation of child self-report of pain or overall well-being.

Clinical usability. The 21–numbered circle VAS is appropriate for clinical use. Specifically, it may be more feasible than a traditional 10-cm horizontal-line VAS as scoring does not require the use of a ruler.

Research usability. The 21–numbered circle VAS is appropriate for research use. It allows for more precise assessment than traditional 10-cm horizontal-line VAS, given that it does not require reproduction of an exact line in printing or photocopying.

E-OUCH ELECTRONIC PAIN DIARY

Description

Purpose. The e-Ouch is a multidimensional electronic diary that uses a real-time data capture approach to measure pain intensity, pain unpleasantness, and pain's interference with features of health-related quality of life (HRQOL) in adolescents with juvenile arthritis (JA). The e-Ouch was originally published in 2006.

Content. A personal digital assistant is programmed to sound an audible alarm to obtain 3 daily pain ratings (upon waking, after school, and before bed). The ratings are designed to evaluate multiple dimensions of pain: 1) sensory discriminant, 2) affective motivational, and 3) cognitive evaluative.

Number of items. The number of items presented is dependent on whether the adolescent endorses having current pain. If the adolescent does not report current pain, the diary does not present detailed pain questions, and advances to questions regarding stiffness and fatigue. Assuming the adolescent reports pain, they will be presented 12 items in the morning, 15 in the afternoon, and 16 in the evening.

Response options/scale. Adolescents move a visual analog scale (VAS) slider anchored at "no pain" and "very much pain" to indicate pain intensity. Pain unpleasantness is evaluated using a VAS slider anchored at "not at all unpleasant" and "very unpleasant." Pain interference is evaluated using several VAS sliders anchored at "doesn't get in the way at all" and "totally gets in the way." Adolescents report on the level of interference in the following activities: "things you do," "how you feel," "walking," "sleeping," "enjoying life," "schoolwork," and "relationships."

Recall period for items. With a few exceptions, items ask adolescents to report how their pain is "right now." The recall period for the remaining items is "today," or "last night" when questioning pain's interference with sleep.

Endorsements. None.

Examples of use. Stinson JN, Petroz GC, Stevens BJ, Feldman BM, Streiner D, McGrath PJ, et al. Working out the kinks: testing the feasibility of an electronic pain diary for adolescents with arthritis. Pain Res Manage 2008;13: 375–82 (21).

Stinson JN, Petroz GC, Tait G, Feldman BM, Streiner D, McGrath PJ, et al. E-Ouch: usability testing of an electronic chronic pain diary for adolescents with arthritis. Clin J Pain 2006;22:295–305 (22).

Practical Application

How to obtain. The e-Ouch was developed by Jennifer N. Stinson at the University of Toronto (Jennifer N. Stinson, PhD, MSc, BScN, Assistant Professor, Lawrence S. Bloomberg Faculty of Nursing, University of Toronto, 155 College Street, Room 276, Toronto, Ontario, M5T 1P8 Canada. E-mail: jennifer.stinson@utoronto.ca).

The e-Ouch is currently being evaluated as part of a study funded by the Childhood Arthritis and Rheumatology Research Alliance. It may be viewed, with permission and a password, online at www.superkidzpain.ca.

Method of administration. Patient completed.

Scoring. The e-Ouch is computer scored. Pain indices are reported across the 3 diary entries. Each VAS is measured in millimeters. Scores for the following VAS are reported: pain intensity, pain unpleasantness, pain interference (i.e., activities, mood, walking, sleep, schoolwork, relationships, and enjoyment of life), stiffness, tiredness, and control over pain. The number of painful joints and number of pain words selected by the adolescent to describe current pain are reported. Missing data are handled by summarizing data over time of day, day of week, and week. E-Ouch diary entries are averaged across each of the 3 time periods separately within weekdays and weekends.

Score interpretation. The range of scores for each VAS is 0–100. Higher scores indicate more pain, unpleasantness, interference, stiffness, tiredness, and control over the pain. Adolescents are able to select each major joint on the body picture. Higher scores indicate a higher number of painful joints. Pain word descriptors are chosen from a list of 30 words, with a range of 0–30.

Respondent burden. Adolescents complete all 3 daily pain ratings in <9 minutes. It is possible that user fatigue may occur, resulting in decreased compliance over time and subsequent missing data. During usability testing, all adolescent participants (n = 20) stated the e-Ouch was "easy to understand" and "self-explanatory."

Administrative burden. The administrative burden associated with the e-Ouch is very low, given automatic electronic data entry and scoring.

Translations/adaptations. None.

Psychometric Information

Method of development. Development of the e-Ouch has occurred over 3 phases: 1) usability (i.e., intuitiveness of the user interface), 2) feasibility (i.e., acceptability and adherence), and 3) psychometric evaluation. Usability testing involved semistructured interviews and observation of e-Ouch use by 10 adolescents with JA (22). Changes were made to the e-Ouch based on this initial feedback, and a second iterative cycle of 10 adolescents was completed. Next, 13 adolescents participated in feasibility testing in 2 cycles (21). Technical difficulties evident in the first cycle were addressed in the second cycle. Psychometric evaluation of the e-Ouch began following refinement of the prototype.

Acceptability. During initial testing, 2 semistructured interviews with 20 adolescents with JA were conducted to evaluate learnability, efficiency, errors, and satisfaction. All of the adolescents stated the e-Ouch was "very easy to learn" and "very easy to use." The majority of adolescents stated the e-Ouch was "quick" to complete. Errors made during the first cycle of testing were corrected and no errors occurred during the second cycle. All of the adolescents were "very satisfied" with the e-Ouch. Authors of a study evaluating the construct validity and feasibility of the e-Ouch reported 22% of data for electronic pain ratings was missing.

Reliability. Not reported.

Validity. Construct validity for the e-Ouch has been established (22). Adolescents with JA completed e-Ouch diary entries 3 times per day for a 2-week period. At the end of the first week, participants used paper VAS ratings to recall their least, average, and worst pain intensity, unpleasantness, and interference rating for the preceding week. At the end of the 2-week period, participants completed the Pediatric Quality of Life (PedsQL) Inventory 4.0, the PedsQL 3.0 Arthritis Module, and the Pain Coping Questionnaire. Statistically significant correlations were found between e-Ouch pain indices and scores from recalled paper VAS ratings. Correlations were positive and ranged from 0.49–0.84. Correlations between e-Ouch pain indices and scores from overall HRQOL, disease-specific HRQOL, and emotion-focused coping ranged from −0.64 to −0.18. Correlations between e-Ouch pain indices and physician-rated disease activity indices were not significant.

Ability to detect change. Evidence suggests that the e-Ouch is responsive to changes in pain intensity, unpleasantness, and interference in adolescents following intra-articular joint injections. Specifically, medium effect sizes have been reported, with a range from 0.52–0.71 (22).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. A significant strength of the e-Ouch is its use in real time, which should minimize recall bias. The e-Ouch is appropriate for research use and may be used to evaluate interventions. E-Ouch diary entries contained no errors. This is a significant strength, especially relative to the high number of errors typical for paper diaries.

Caveats and cautions. Analysis of e-Ouch data requires challenging statistical analyses that may be unfamiliar to many statisticians. Prompting participants to complete diary entries at prespecified reporting times (i.e., a signal-contingent approach) might result in reporting bias due to prespecified reporting times. Further research is needed to examine the effect that self-monitoring may have on pain reports over time in adolescents using the e-Ouch.

Clinical usability. Consistent use of the e-Ouch in a clinical setting will be more feasible if it can be adapted for electronic medical record format. Likewise, the e-Ouch may be more usable if developed as an application for smart phones (e.g., iPhone) or other personal electronic devices (e.g., iPod).

Research usability. Initial reports of the psychometric properties of the e-Ouch are promising and provide support for further evaluation. The e-Ouch is feasible for some research use, but will be limited by the resources available to the researcher.

OUCHER

Description

Purpose. The Oucher is a single-item measure of pain intensity in children ages 3–12 years. The Oucher has been used extensively in numerous pediatric pain populations. It has been used in 1 published study of children with juvenile arthritis. The Oucher was developed by Judith E. Beyer in 1980. The first published study using the Oucher was in 1992. The metric used for the Oucher was changed from 0–100 to 0–10 in 2000 to be consistent with other visual analog scale (VAS) pain measures (24). A recent review of evidence-based pediatric pain measures rated the Oucher as "well-established" (25).

Content. The Oucher contains 2 separate scales: a numerical scale and a photographic scale. Only 1 scale is used with any given child. The Oucher provides a measure of current pain intensity.

Number of items. The Oucher contains 1 item measuring child self-reported pain intensity.

Response options/scale. The child selects a number (i.e., 1–10) or photograph (i.e., 1 of 6) corresponding to the amount of "hurt" they have.

Recall period for items. Children are asked, "How much hurt do you have right now?"

Endorsements. None.

Examples of use. Schanberg LE, Lefebvre JC, Keefe FJ, Kredich DW, Gil KM. Pain coping and the pain experience in children with juvenile chronic arthritis. Pain 1997;73: 181–9 (26).

Practical Application

How to obtain. The Oucher posters are not currently being produced, but remaining inventory is available for purchase. Posters of the Caucasian, African American, and Hispanic versions of the Oucher may be purchased for $2 each at the following address: Pain Associates in Nursing, PO Box 411714, Kansas City, MO 64141. E-mail address: info@oucher.org. All versions of the Oucher may be downloaded for free from http://www.oucher.org/.

Method of administration. Patient self-report.

Scoring. The Oucher does not require computer scoring. If the photographic scale is used, the photograph selection must be converted to an ordinate scale ranging from 0–5: the bottom photograph is scored as 0 and the top photograph is scored as 5. If the numerical scale is used, the child's selection can be recorded directly; no conversions are necessary.

Score interpretation. The score range for the numerical scale of the Oucher is 0 ("no hurt") to 10 ("biggest hurt of all"). The photographic scale is anchored by 6 photographs displaying varying degrees of discomfort.

Respondent burden. Following initial training on how to use the Oucher, ~15 seconds is required to complete the measure.

Administrative burden. Approximately 3–4 minutes are required to train a child to use the Oucher. Prior to administering the Oucher, children are asked to complete a series of cognitive tasks to determine which scale is appropriate to administer. Children use the numerical scale if they can count to 100 by ones or tens and they can identify which of any 2 numbers is larger.

Translations/adaptations. There are 5 versions of the Oucher currently available: 1) white or Caucasian, 2) black or African American, 3) Hispanic, 4) First Nations (boy and girl), and 5) Asian (boy and girl). Psychometric evaluations of the various versions of the Oucher have generally been conducted with children in the ethnic group depicted in the Oucher photographs.

Psychometric Information

Method of development. The Oucher was developed in line with several recommendations for measures of pediatric pain: 1) required only simple instructions, 2) appropriate for children ages 3–12 years, 3) direct pain or discomfort cues provided, and 4) verbal communication not required (26). Photographs of a child's face while in pain or discomfort were chosen as the direct cue. The photographs were arranged from a neutral expression to one in which the child's face was distorted in pain or discomfort.

Acceptability. Not available.

Reliability. Test–retest reliability for the Oucher has been evaluated indirectly. One study used the Charleston Pediatric Pain Pictures to present hypothetical pain stimuli to 50 nonpatient preschoolers ages 3–6 years. The pictures were accompanied by a brief vignette and depicted scenes commonly experienced by preschool children. Each picture was previously rated by 6 experienced child clinicians as representing no pain (e.g., looking at a picture book at home), low pain (e.g., having an adhesive bandage removed), moderate pain (e.g., stubbing toe on sidewalk), and high pain (e.g., burning hand on stove at home). For each of 17 pictures, participants were instructed to use 3 different measures of pain intensity (i.e., Oucher, Pain Thermometer, and Faces Scale) to indicate "how much hurt you would have" in each picture. Thirty-six of the children rated the pictures again 1 week later. The average test–retest correlation of individual items rated using the Oucher was 0.43, with a range of 0.11–0.83 (28).

Validity. Numerous studies have examined the validity of the Oucher. Content validity was established in a sample of 78 children ages 3–7 years. The children arranged the 6 photographs of the original Oucher according to their perception of least to most hurt. Agreement, reported as Kendall's coefficient of concordance, was 0.73 (29). Construct validity was evaluated in the study described above in Reliability. Intercorrelations between the Oucher and the Pain Thermometer ranged from 0.62–0.86. Intercorrelations between the Oucher and the Faces Scale ranged from 0.70–0.88 (28).

Ability to detect change. Evidence of the responsiveness of the Oucher was provided by a study of 25 children between the ages of 3.0 and 12.4 years hospitalized for traumatic injuries or surgery. Participants used 3 measures of pain intensity (i.e., the Oucher, the Poker Chip Tool, and a VAS) to provide postoperative pain ratings. Pain ratings occurred within 30 minutes before receiving analgesic medication and at four 1-hour intervals after receiving medication. The mean preanalgesic pain score on the numerical scale of the Oucher was 70, with a range of 30–100. Mean postanalgesic pain scores over the 4-hour period ranged from 29.9–41.3. Paired-samples *t*-tests demonstrated that postanalgesic pain scores were significantly lower ($P < 0.01$) at each time interval. The median preanalgesic pain score on the photographic scale of the Oucher ($n = 7$) was 2, with a range of 2–5. Postanalgesic pain scores ranged from 0–3, with a median of 1. Mean and median postanalgesic scores for each time period were lower than preanalgesic scores for all participants (30).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The Oucher is the only pediatric pain tool that includes color photographs of real children who are in pain or discomfort. Other scales use simple line drawings to depict faces. It is appropriate for clinical use. The Oucher attempts to directly address ethnic differences in assessment of pediatric pain.

Caveats and cautions. Additional evidence is required to demonstrate the validity and appropriateness of the use of the various versions of the Oucher as culturally-specific measures of pediatric pain.

Clinical usability. The Oucher is appropriate for clinical use. It is likely to be most useful as a basic measure of pain intensity that may be routinely administered in an effort to monitor pediatric pain.

Research usability. Current psychometric research supports the use of the Oucher for research purposes. More studies are needed to validate the different ethnic versions of the Oucher.

PAIN BEHAVIOR OBSERVATION METHOD

Description

Purpose. The Pain Behavior Observation Method measures pain behaviors in children with juvenile arthritis (JA). It is the only observational pain behavior measure validated for use with this population. It was originally published in 1995.

Content. The Pain Behavior Observation Method uses an interval sampling method to measure the frequency of 6 pain behaviors: 1) guarding, 2) bracing, 3) active rubbing, 4) rigidity, 5) single flexing, and 6) multiple flexing.

Number of items. Six pain behaviors are coded.

Response options/scale. Observers code whether each of the 6 pain behaviors occurs during a total of twenty 30-second intervals.

Recall period for items. Not applicable.

Endorsements. None.

Examples of use. Jaworski TM, Bradley LA, Heck LW, Roca A, Alarcon GS. Development of an observation method for assessing pain behaviors in children with juvenile rheumatoid arthritis. Arthritis Rheum 1995;38: 1142–51 (31).

Practical Application

How to obtain. The Pain Behavior Observation Method was developed by Theresa M. Jaworski while at the University of Alabama at Birmingham (Theresa M. Jaworski, PhD, Licensed Psychologist, 6507 Transit Road, Suite B, East Amherst, NY 14051. E-mail: tmj3@buffalo.edu).

Method of administration. Clinician completed.

Scoring. Children perform a standardized sequence of behaviors (two 1-minute sitting periods, two 1-minute standing periods, two 1-minute reclining periods, and four 1-minute walking periods). A trained observer views the videotape and codes pain behaviors using an interval-sampling method for a total of twenty 30-second intervals (with a 20-second observation phase followed by a 10-second recording phase for each 30-second interval). Following this method, observed behaviors are only coded as occurring once during any 20-second observation phase.

Score interpretation. The range of scores is 0–20 for each of the 6 pain behaviors. The range of total pain behavior scores is 0–120. Higher scores indicate greater number of pain behaviors.

Respondent burden. The Pain Behavior Observation Method requires ~10 minutes to complete.

Administrative burden. The Pain Behavior Observation Method requires ~10 minutes to administer the standardized sequence of behaviors and 10–20 minutes to hand score. Extensive training is required to administer and score the Pain Behavior Observation Method.

Translations/adaptations. None.

Psychometric Information

Method of development. The Pain Behavior Observation Method was adapted from an observational method developed by McDaniel and colleagues for use with adults with rheumatoid arthritis (32). Six children were videotaped while completing a sequence of activities (e.g., sitting and walking). These videotapes were reviewed by 4 chronic pain experts to identify and operationally define frequent pain behaviors specific for children and adolescents with JA.

Acceptability. The physical maneuvers required to conduct the Pain Behavior Observation method are doable for children with JA.

Reliability. The percentage of overall interrater agreement is 90–95%. The percentage of effective agreement (i.e., occurrences only) is 63–87%. Kappa coefficients range from 0.53–0.79 (31).

Validity. Correlations between pairs of individual pain behaviors are generally not significant. The amount of variance shared by all possible pairs of behaviors ranged from 13–25%. The total pain behavior score is significantly correlated with functional disability (r = 0.64, P = 0.0001), but not significantly correlated with self-reports of depression. The total pain behavior score is significantly correlated with children and parent visual analog scale ratings of pain (r = 0.50, P = 0.005 and r = 0.48, P = 0.007, respectively) (31).

Ability to detect change. Not reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. This method is likely to be particularly useful to assess pain in those children who are not able to provide reliable and valid reports of pain (i.e., younger children or those with cognitive deficits). The Pain Behavior Observation Method may provide useful supplemental data to supplement other measures of pain.

Caveats and cautions. This measure would not be appropriate for use with children who have difficulty ambulating. To date, there is only 1 published study evaluating the Pain Behavior Observation Method with 30 participants. More research is needed to further evaluate the usefulness of the Pain Behavior Observation Method.

Clinical usability. This method has a number of weaknesses making it not feasible for clinical use. Both administrative and respondent burden are likely to limit clinical use. Extensive training is required to use the Pain Behavior Observation Method. Patients must be able to perform a standardized sequence of behaviors on videotape. Behaviors associated with chronic pain (e.g., joint pain) are difficult to observe reliably due to associated pain habituation.

Research usability. Current psychometric evaluation supports continued research use. Administrative and respondent burden may limit research use to those projects with the resources required to use the Pain Behavior Observation Method.

PEDIATRIC PAIN ASSESSMENT TOOL (PPAT)

Description

Purpose. The PPAT is a multidimensional measure of pediatric pain intensity. The PPAT has been used with school-aged children to measure pain associated with juvenile arthritis (JA), cancer, and surgical operations. It was originally developed in The Netherlands and published in 1990. A recent review of evidence-based pediatric pain measures rated the PPAT as "approaching well-established" (24).

Content. The PPAT assesses the sensory, affective, and evaluative domains of pediatric pain.

Number of items. 32 word descriptors and a 10-cm visual analog scale (VAS) for current and worst pain intensity.

Response options/scale. A 10-cm VAS with 1-cm gradations (where 1 = "I have no pain" and 10 = "I have very severe pain") measures the child's present and worst pain. The child chooses from a list of 32 word descriptors of various aspects of pain.

Recall period for items. Word descriptors, VAS for current pain intensity, and VAS for worst pain intensity "this week."

Endorsements. A Society of Pediatric Psychology task force on evidence-based measures of pain in children identified the PPAT as "approaching well-established."

Examples of use. Abu-Saad HH, Uiterwijk M. Pain in children with juvenile rheumatoid arthritis: a descriptive study. Pediatr Res 1995;38:194–7 (33).

Practical Application

How to obtain. The PPAT was developed by Huda Huijer Abu-Saad while at the University of Limburg, The Netherlands (Huda Huijer Abu-Saad, RN, PhD, FEANS, Professor of Nursing Science, Director, School of Nursing, Faculty of Medicine, American University of Beirut. E-mail: Huda.Huijer@aub.edu.lb).

Method of administration. Patient self-report, parent report, physician report, nurse report, and interviewer administered.

Scoring. Pain intensity is scored by measuring the 10-cm scale with a ruler. The word descriptors may be scored 2 ways. The first is to count the number of word descriptors the child chooses to describe their pain. The second is to compute an average pain intensity score from the intensity scores for each of the selected word descriptors.

Respondent burden. The PPAT requires ~5–10 minutes to complete. School-aged children report no difficulty understanding the terms used.

Administrative burden. The PPAT requires ~5–10 minutes to administer and score.

Translations/adaptations. The PPAT was originally developed for use in The Netherlands. It has been administered to Arab-American (34) and Jordanian (35) children.

Psychometric Information

Method of development. The McGill Pain Questionnaire (36) and Pediatric Pain Questionnaire (37) served as models for the development of the PPAT. Ten children ages 9–15 years hospitalized for surgical procedures were asked to describe their pain (38). Their responses were recorded verbatim. A similar procedure was later conducted with 355 healthy children ages 7–17 years (39).

Acceptability. In the 1 study using the PPAT with a JA population, the authors reported that children ages 7–16 years had no difficulty describing their pain using the PPAT (33).

Reliability. Interrater agreement correlations between child, parent, and physician VAS pain intensity ratings ranged from 0.32–0.77. Agreement between parent and physician VAS current pain ratings were not significant ($r = 0.10$) (33). A Cronbach's alpha of 0.83 was reported for the PPAT (39).

Validity. Content validity has been established (39). Evidence of construct validity has been provided in a factor analysis (39) and in a study investigating postoperative pain reports of 105 children ages 5–15 years before and after analgesic administration (17). This study also provided evidence of convergent validity (17). Participants rated postoperative pain using the 10-cm scale and word descriptors from the PPAT, the Oucher, and a 100-mm VAS. Correlations between the 10-cm scale of the PPAT, PPAT word descriptors, the Oucher, and the 100-mm VAS ranged from 0.88–0.98. Correlations between the number of word descriptors and the 10-cm scale of the PPAT, the Oucher, and the 100-mm VAS ranged from 0.47–0.81. Correlations between the word descriptors and the 10-cm scale of the PPAT, the Oucher, and the 100-mm VAS ranged from 0.02–0.67. Evidence of divergent validity was provided by correlating postoperative pain reports using multiple measures and a scale measuring fear in children (i.e., The Child Medical Fear Scale). Correlations ranged from 0.14–0.26 (17).

Ability to detect change. Not reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PPAT is based on comprehensive theories of pain. It is easy to score and administer and has adequate reliability and validity.

Caveats and cautions. The PPAT has been used in 1 study with 33 children with JA. The ability of the PPAT to detect change has not been established.

Clinical usability. The PPAT is feasible for clinical use, but more studies are needed with children and adolescents with JA.

Research usability. The PPAT is appropriate for research use.

PEDIATRIC PAIN QUESTIONNAIRE (PPQ)

Description

Purpose. The PPQ is a measure of pain intensity and location. It has primarily been applied to children with sickle cell disease and juvenile arthritis. It was originally published in 1987. A recent review of evidence-based pediatric pain measures rated the PPQ as "well-established" (25).

Content. The PPQ assesses the sensory, affective, and evaluative domains of pediatric pain.

Number of items. 6 items.

Response options/scale. The sensory domain is assessed with child visual analog scale (VAS) ratings of pain intensity and a body outline to describe the location of pain. Having the child choose words that describe their current pain or how they feel when in pain assesses affective and evaluative domains.

Recall period for items. VAS for current and worst pain "this week" and body outline for current pain.

Endorsements. A Society of Pediatric Psychology task force on evidence-based measures of pain in children identified the PPQ as "well-established."

Examples of use. Varni JW, Thompson KL, Hanson V. The Varni/Thompson Pediatric Pain Questionnaire. I. Chronic musculoskeletal pain in juvenile rheumatoid arthritis. Pain 1987;28:27–38 (37).

Practical Application

How to obtain. The PPQ was developed by James W. Varni. It may be obtained from www.pedsql.org.

Method of administration. Patient self-report, parent report, or interviewer administered.

Scoring. Pain intensity is scored by measuring the VAS with a ruler. The body outline is used to score the number of body sites with current pain and intensity.

Score interpretation. Range of pain intensity scores using the VAS is 0 (no pain) to 100 (severe pain). Range of pain intensity scores using the body outline is 0 (none) to 3 (severe).

Respondent burden. The PPQ requires ~10–15 minutes to complete.

Administrative burden. The PPQ requires <5 minutes to hand score.

Translations/adaptations. The PPQ has been translated into Danish, Norwegian, Portuguese, Spanish, Swedish, and French.

Psychometric Information

Method of development. The PPQ was adapted from the McGill Pain Questionnaire developed by Melzack for use with adults (36). Pediatric psychologists and rheumatologists reviewed items for content appropriateness and feasibility of use with children and adolescents.

Acceptability. Content was deemed developmentally appropriate for children ages 4–16 years.

Reliability. Test–retest reliability correlations for 1-week, 3-week, and 6-month intervals ranged from 0.29–0.41 (2). Interrater agreement correlations between child, parent, nurse, and physician VAS pain intensity ratings ranged from 0.40–0.85 (40).

Validity. Construct validity has been established. Evidence of convergent validity of the PPQ with disease status ranged from 0.27–0.68, and ranged from 0.06–0.45 with psychological functioning (2).

Ability to detect change. The PPQ is commonly used in pain treatment studies to document changes in pain intensity following intervention.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PPQ is a widely disseminated measure of pediatric pain. It assesses pediatric pain using a developmentally appropriate format. The PPQ is appropriate for both clinical and research use.

Caveats and cautions. There is currently no comprehensive manual with detailed instructions for use of the PPQ.

Clinical usability. The body outline used in the PPQ is particularly useful to assess pediatric pain in a clinical setting.

Research usability. The words generated by children and those selected from a list of sensory, affective, and evaluative words is particularly useful for research purposes.

DISCUSSION

Compared to our original review article of pediatric pain measures for juvenile arthritis (JA) (41), 2 of 3 made it into this current review: the Pain Behavior Observation Method and the Pediatric Pain Questionnaire (PPQ). We dropped the Pain Coping Questionnaire from the previous review because it was not a measure of pain per se. In this review, we added 3 additional measures: the 21–numbered circle visual analog scale (VAS), the e-Ouch Electronic Pain Diary, and the Oucher, which have all been used with children and adolescents with JA. Future research needs to validate these measures in other pediatric rheumatic conditions.

The PPQ has been most widely used and is considered a "well-established" instrument by empirical standards for measuring pain (24). The 21–numbered circle VAS looks very promising and includes parental ratings of their child's global well-being and physician ratings of overall disease activity, which are criteria endorsed by the American College of Rheumatology for demonstrating responsiveness to treatments for JA. The downside to the 21–numbered circle VAS is that there is no child self-report version. The Oucher has been used to assess pediatric acute and chronic pain since being developed by Judith Beyer in 1980, and is a well-established instrument with different ethnic versions and a picture versus a numerical VAS that can be used by younger children. The e-Ouch has very promising VAS and a body outline. Electronic measures are going to be the wave of the future as we move into paperless electronic medical records (EMRs). We need to develop e-versions of pain measures for EMRs as well as web-based programs for children and adolescents with JA and other rheumatic diseases.

AUTHOR CONTRIBUTIONS

Both authors were involved in drafting the article or revising it critically for important intellectual content, and both authors approved the final version to be published.

REFERENCES

1. Rapoff MA, Lindsley CB. Pain. In: Cassidy JT, Petty RE, Laxer RM, Lindsley CB, editors. Textbook of pediatric rheumatology. 6th ed. Philadelphia: Elsevier; 2011. p. 192–7.
2. Gragg RA, Rapoff MA, Danovsky MB, Lindsley CB, Varni JW, Waldron SA, et al. Assessing chronic musculoskeletal pain associated with rheumatic disease: further validation of the Pediatric Pain Questionnaire. J Pediatr Psychol 1996;21:237–50.
3. Hagglund KJ, Schopp LM, Alberts KR, Cassidy JT, Frank RG. Predicting pain among children with juvenile rheumatoid arthritis. Arthritis Care Res 1995;8:36–42.
4. Ilowite NT, Walco GA, Pochaczevsky R. Assessment of pain in patients with juvenile rheumatoid arthritis: relation between pain intensity and degree of joint inflammation. Ann Rheum Dis 1992;51:343–6.
5. Thompson KL, Varni JW, Hanson V. Comprehensive assessment of pain in juvenile rheumatoid arthritis: an empirical model. J Pediatr Psychol 1987;12:241–55.
6. Varni JW, Rapoff MA, Waldron SA, Gragg RA, Bernstein BH, Lindsley CB. Chronic pain and emotional distress in children and adolescents. J Dev Behav Pediatr 1996;17:154–61.
7. Ross CK, Lavigne JV, Hayford JR, Dyer AR, Pachman LM. Validity of reported pain as a measure of clinical state in juvenile rheumatoid arthritis. Ann Rheum Dis 1989;48:817–9.
8. Schanberg LE, Lefebvre JC, Keefe FJ, Kredich DW, Gil KM. Pain coping and the pain experiences in children with juvenile chronic arthritis. Pain 1997;73:181–9.
9. Benestad B, Vinje O, Veierod MB, Vandvik IH. Quantitative and qualitative assessments of pain in children with juvenile chronic arthritis based on the Norwegian version of the Pediatric Pain Questionnaire. Scand J Rheumatol 1996;25:293–9.
10. Schanberg LE, Anthony KK, Gil KM, Maurin EC. Daily pain and symptoms in children with polyarticular arthritis. Arthritis Rheum 2003;48:1390–7.
11. Lovell DJ, Walco GW. Pain associated with juvenile rheumatoid arthritis. Pediatr Clin North Am 1989;36:1015–27.
12. Peterson LS, Mason T, Nelson AM, O'Fallon WM, Gabriel SE. Psychosocial outcomes and health status of adults who have had juvenile rheumatoid arthritis. Arthritis Rheum 1997;40:2235–40.

13. Pincus T, Bergman M, Sokka T, Roth J, Swearingen C, Yazici Y. Visual analog scales in formats other than a 10 centimeter horizontal line to assess pain and other clinical data. J Rheumatol 2008;35:1550−8.
14. Stinson JN, Kavanagh T, Yamada J, Gill N, Stevens B. Systematic review of the psychometric properties, interpretability and feasibility of self-report pain intensity measures for use in clinical trials in children and adolescents. Pain 2006;125:143–57.
15. Von Baeyer CL, Spagrud LJ. Systematic review of observational (behavioral) measures of pain for children and adolescents aged 3 to 18 years. Pain 2007;127:140−50.
16. Uziel Y, Chapnick G, Rothschild M, Tauber T, Press J, Harel L, et al. Nitrous oxide sedation for intra-articular injection in juvenile idiopathic arthritis. J Pediatr Rheumatol Online J 2008;6:1.
17. Abu-Saad HH, Pool H, Tulkens B. Further validity testing of the Abu-Saad Paediatric Pain Assessment Tool. J Adv Nurs 1994;19:1063–71.
18. McGrath PJ, Walco G, Turk DC, Dworkin RH, Brown MT, Davidson K, et al. Core outcome domains and measures for pediatric acute and chronic/recurrent pain clinical trials: PedIMMPACT recommendations. J Pain 2008;9:771−83.
19. Filocamo G, Davi S, Pistorio A, Bertamino M, Ruperto N, Lattanzi B, et al. Evaluation of 21-numbered circle and 10-centimeter horizontal line visual analog scales for physician and parent subjective ratings in juvenile idiopathic arthritis. J Rheumatol 2010;37:1534−41.
20. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. Arthritis Rheum 1997;40:1202–9.
21. Stinson JN, Petroz GC, Stevens BJ, Feldman BM, Streiner D, McGrath PJ, et al. Working out the kinks: testing the feasibility of an electronic pain diary for adolescents with arthritis. Pain Res Manage 2008;13: 375−82.
22. Stinson JN, Petroz GC, Tait G, Feldman BM, Streiner D, McGrath PJ, et al. E-Ouch: usability testing of an electronic chronic pain diary for adolescents with arthritis. Clin J Pain 2006;22:295–305.
23. Stinson JN, Stevens BJ, Feldman BM, Streiner D, McGrath PJ, Dupuis A, et al. Construct validity of a multidimensional electronic pain diary for adolescents with arthritis. Pain 2008;136:281–92.
24. Von Baeyer CL, Hicks CL. Support for a common metric for pediatric pain intensity scales. Pain Res Manage 2000;4:157−60.
25. Cohen LL, Lemanek K, Blount RL, Dahlquist LM, Lim CS, Palermo TM, et al. Evidence-based assessment of pediatric pain. J Pediatr Psychol 2008;33:939−55.
26. Schanberg LE, Lefebvre JC, Keefe FJ, Kredich DW, Gil KM. Pain coping and the pain experience in children with juvenile chronic arthritis. Pain 1997;73:181–9.
27. Beyer J, Aradine C. Patterns of pediatric pain intensity: a methodological investigation of a self-report scale. Clin J Pain 1987;3:130−41.
28. Belter R, McIntosh J, Finch A, Saylor C. Preschoolers' ability to differentiate levels of pain: relative efficacy of three self-report measures. J Clin Child Psychol 1988;17:329−35.
29. Beyer J, Aradine C. Content validity of an instrument to measure young children's perceptions of the intensity of their pain. J Pediatr Nurs 1986;1:386−95.
30. Aradine CR, Beyer JE, Tompkins JM. Children's pain perception before and after analgesia: a study of instrument construct validity and related issues. J Pediatr Nurs 1988;3:11–23.
31. Jaworski TM, Bradley LA, Heck LW, Roca A, Alarcon GS. Development of an observation method for assessing pain behaviors in children with juvenile rheumatoid arthritis. Arthritis Rheum 1995;38:1142–51.
32. McDaniel LK, Anderson KO, Bradley LA, Young LD, Turner RA, Agudelo CA, et al. Development of an observational method for assessing pain behavior in rheumatoid arthritis patients. Pain 1986;24:165–84.
33. Abu-Saad HH, Uiterwijk M. Pain in children with juvenile rheumatoid arthritis: a descriptive study. Pediatr Res 1995;38:194−7.
34. Abu-Saad H. Cultural components of pain: the Arab-American child. Issues Compr Pediatr Nurs 1984;7:91–9.
35. Gharaibeh M, Abu-Saad H. Cultural validation of pediatric pain assessment tools: Jordanian perspective. J Multicult Nurs Health 2002;13: 12−8.
36. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. Pain 1975;1:277–99.
37. Varni JW, Thompson KL, Hanson V. The Varni/Thompson Pediatric Pain Questionnaire. I. Chronic musculoskeletal pain in juvenile rheumatoid arthritis. Pain 1987;28:27–38.
38. Abu-Saad HH. Assessing children's responses to pain. Pain 1984;19: 163–71.
39. Abu-Saad HH, Kroonen E, Halfens R. On the development of a multidimensional Dutch pain assessment tool for children. Pain 1990;43: 249−56.
40. Thompson KL, Varni JW, Hanson V. Comprehensive assessment of pain in juvenile rheumatoid arthritis: an empirical model. J Pediatr Psychol 1987;12:241–55.
41. Rapoff MA. Pediatric measures of pain: the Pain Behavior Observation Method, Pain Coping Questionnaire (PCQ), and Pediatric Pain Questionnaire (PPQ). Arthritis Rheum 2003;49 Suppl:S90−5.

Summary Table of Pediatric Pain Measures*

Scale	Purpose/ content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
21–numbered circle VAS	A VAS with 21 circles in 0.5-unit increments measuring overall well-being, pain intensity, and overall disease activity	Parent and physician report	~5–10 seconds to complete	7.4 seconds to hand score	VAS, 0 (very well) to 10 (very poorly), 0 (no pain) to 10 (very severe pain), 0 (no activity) to 10 (maximum activity)	NR	Construct validity established	Responsiveness to change over time and minimum clinically significant important difference established	Easy to administer and score, easy to reproduce, appropriate for clinical use	No child self-report, ceiling effects of disease activity
E-Ouch electronic pain diary	Electronic pain diary measure validated with adolescents who have JA Pain intensity, pain unpleasantness, and pain interference	Patient self-report	3 daily pain ratings requiring <9 minutes to complete	Computer scored	100-mm VAS for pain intensity, pain unpleasantness, pain interference, stiffness, and fatigue; body picture (major joints); pain word list (30 pain word descriptors)	NR	Construct validity established	Detects changes in pain ratings in adolescents undergoing joint injections	Used in real time, few entry errors, appropriate for research use, feasible in a home setting	Potential reactive effects and reporting bias, challenging data analyses
Oucher	Self-report measure of pain intensity	Patient self-report	~15 seconds to complete once trained to use the Oucher	3–4 minutes to train child to use the Oucher	Numerical scale: 0 (no hurt) to 10 (biggest hurt of all); photographic scale: 0 (no hurt) to 10 (biggest hurt of all)	Test–retest range 0.11–0.83	Content and construct validity established	Detects changes in postoperative pain ratings in children	Appropriate for clinical use, uses color photographs	Further validation of different ethnic versions is needed
Pain Behavior Observation Method	Observational pain behavior measure validated with children who have JA 6 pain behaviors (guarding, bracing, active rubbing, rigidity, single flexing, and multiple flexing)	Clinician completed	~10 minutes to complete	~10 minutes to administer, 10–20 minutes to hand score	Score range 0–20 for each of the 6 pain behaviors; total pain behavior score range 0–120, where higher scores represent greater number of pain behaviors	Interrater agreement: acceptable	Concurrent validity established	NR	Appropriate for research use, useful with children unable to provide reliable and valid pain reports	Requires extensive training, not feasible for clinical use, limited research
PPAT	10-cm scale of pain intensity and 32 word descriptors of the sensory and affective/ evaluative domains of pain	Patient self-report, parent, nurse, and physician proxy report	5–10 minutes to complete	5–10 minutes to administer and score	10-cm scale: 0 (no hurt) to 10 (severe hurt); pain word list: number of word descriptors and word descriptor intensity	Interrater agreement: acceptable Cronbach's $\alpha = 0.83$	Content, construct, convergent, and divergent validity established	NR	Based on comprehensive pain theories, easy to score and administer, appropriate for clinical and research use	Further use in JA population is needed
PPQ	A measure of pain intensity (VAS) and location (body outline) and the sensory, affective, and evaluative qualities of pain (words describing pain), appropriate for children and adolescents	Patient self-report or interviewer administered	10–15 minutes to complete	5 minutes to hand score	Pain intensity range using VAS 0 (no pain) to 100 (severe pain); body outline intensity range 0 (none) to 3 (severe)	Moderate stability Interrater agreement: acceptable	Concurrent validity established	Used in pain treatment studies to record pre- to posttreatment changes in pain intensity	Widely distributed, appropriate for both clinical and research use	No inclusive manual with detailed instructions

* When information was not available in the published literature, we contacted authors via e-mail to attempt to retrieve information, most notably for ease of administration and scoring. VAS = visual analog scale; NR = not reported; JA = juvenile arthritis; PPAT = Pediatric Pain Assessment Tool; PPQ = Pediatric Pain Questionnaire.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S4–S13
DOI 10.1002/acr.20620

Measures of Functional Status and Quality of Life in Rheumatoid Arthritis

Health Assessment Questionnaire Disability Index (HAQ), Modified Health Assessment Questionnaire (MHAQ), Multidimensional Health Assessment Questionnaire (MDHAQ), Health Assessment Questionnaire II (HAQ-II), Improved Health Assessment Questionnaire (Improved HAQ), and Rheumatoid Arthritis Quality of Life (RAQoL)

LEANN MASKA,[1] JACLYN ANDERSON,[2] AND KALEB MICHAUD[3]

INTRODUCTION

Rheumatoid arthritis (RA) is one of the few diseases where subjective patient and physician measures are the best predictors of treatment response and future health outcomes. Arguably, the most important of these is the American College of Rheumatology core measure of function. Developed in 1978, the original Health Assessment Questionnaire Disability Index (HAQ) remains the gold standard for measuring functional status in RA (1). However, its length at 41 questions and relatively complex scoring can make clinical use difficult. We summarize the HAQ and the most common measures developed from it including the Modified HAQ, Multidimensional HAQ, HAQ-II, and Improved HAQ (2–5). Although it is not a primary measure of function, we also review the Rheumatoid Arthritis Quality of Life scale as it is the lone RA-specific quality of life measure and, whether correct or not, functional measures are often used as a substitute for quality of life in RA (6). We did not include several measures that have less recent published use and may be promising for future studies, including the visual analog scale for function (7) and the Patient-Reported Outcomes Measurement Information System computerized adaptive test for function (8).

HEALTH ASSESSMENT QUESTIONNAIRE DISABILITY INDEX (HAQ)

Description

Purpose. Sometimes referred to as the HAQ DI, original, or "legacy" HAQ, the HAQ was developed to assess functional status in adults with arthritis, but is now commonly used among many disciplines (9). Originally developed for use in patients with rheumatoid arthritis (RA) and osteoarthritis, the HAQ has had application in both adults and children within a wider range of rheumatologic conditions including juvenile idiopathic arthritis, systemic lupus erythematosus, systemic sclerosis, ankylosing spondylitis, fibromyalgia, and psoriatic arthritis (9). Additional populations have included human immunodeficiency virus/acquired immunodeficiency syndrome patients and disabled workers (10,11). The measure has also been used to study normal aging as well as for population-based studies (9). Although physical function is only one of several domains determining health-related quality of life, its importance in RA, as well as its prevalence of use, has led to HAQ scores being used to estimate health utilities with a variety of derivations (12–18).

Content. Eight categories, reviewing a total of 20 specific functions evaluate patient difficulty with activities of daily living over the past week. Categories include dressing and grooming, arising, eating, walking, hygiene, reaching, gripping, and errands and chores. Also identified are specific aids or devices utilized for assistance, as well as help needed from another person (aids/help).

Number of items. There are 41 total items: 20 4-point Likert-scale questions assessing specific activities of daily living, 13 additional questions assessing use of assistive devices, and 8 additional questions assessing help received from another. Computation of an Alternative Disability Index (or Alternative HAQ score) is made possible by not taking into account questions regarding the use of aids/help (4,19,20).

[1]Leann Maska, MD: University of Nebraska Medical Center, Omaha, Nebraska; [2]Jaclyn Anderson, DO, MS: Abbott Laboratories, Abbott Park, Illinois; [3]Kaleb Michaud, PhD: University of Nebraska Medical Center, Omaha, Nebraska, and National Data Bank for Rheumatic Diseases, Wichita, Kansas.

Dr. Anderson owns stock and/or holds stock options in Abbott Laboratories.

Address correspondence to Kaleb Michaud, PhD, 986270 Nebraska Medical Center, Omaha, NE 68198-6270. E-mail: kmichaud@unmc.edu.

Submitted for publication February 7, 2011; accepted in revised form May 10, 2011.

Responses options/scale. Twenty specific activities are assessed on a 4-point Likert scale where 0 = without difficulty, 1 = with some difficulty, 2 = with much difficulty, and 3 = unable to do. The 20 activities are grouped into 8 functional categories with each category given a single score equal to the maximum value of their component activities (0, 1, 2, or 3).

Recall period for items. One week.

Practical Application

How to obtain. Developer contact information: James F. Fries, MD, Division of Immunology and Rheumatology, Stanford University Medical Center, 1000 Welch Road, Suite 203, Palo Alto, CA 94304-1808. The HAQ may be used free of charge and is available at http://aramis.stanford.edu/HAQ.html (4).

Method of administration. Most frequently self-administered, but may also be administered by an in-person or telephone interviewer.

Scoring. There are 3 steps to scoring the HAQ (with aids/help): 1) identify the highest subcategory score from each of the 8 categories. Adjust for use of aids/help by increasing the category score from 0 or 1 to a 2 if use of aids/help for that category (utilize table of companion aids/help for HAQ categories). If the category score is already a 2 or 3, no adjustment is made; 2) sum the category scores; and 3) divide the final sum by the number of categories answered to obtain the final HAQ score rounded to the nearest value evenly divisible by 0.125. Requires a minimum of 6 categories answered; if less, do not score.

Score interpretation. Total score is between 0–3.0, in 0.125 increments. Increasing scores indicate worse functioning with 0 indicating no functional impairment and 3 indicating complete impairment. Analyzing nearly 9,000 patients with RA, those identified as independent had a mean ± SD HAQ score of 0.38 ± 0.45. Similarly, those very satisfied with their health had a score of 0.42 ± 0.53 (21). In a non-RA population-based cohort, those with high health risks had significantly higher scores than those with low health risks (HAQ score 1.02 versus 0.49, respectively, $P < 0.001$) (22). Recent studies show that after an immediate rise in HAQ at RA onset, mean HAQ scores increase slowly over time (0.01–0.016 units per year) similar to the general population and are affected by treatment and comorbid conditions (23–25).

Respondent burden. Time to administer/complete is ~5 minutes.

Administrative burden. Time to score is <2 minutes. The median time to score was measured at 24 seconds (20). Training to score should include basic familiarity with the method of scoring and use of aids/help table.

Translations/adaptations. Translations using more than 60 languages and dialects have been performed, although the HAQ was originally developed and validated for English-speaking populations. A list of translations is presented in a 2003 review (9).

Psychometric Information

Reliability. Correlations of test–retest range from 0.87–0.99 (1,21). For each of the 8 subcategories, Spearman's rank correlation has been shown as dressing 0.60, arising 0.82, eating 0.85, walking 0.83, hygiene 0.56, reach 0.80, grip 0.64, and index 0.88. Good repeatability has been demonstrated in RA patients with intraclass correlation coefficient >0.95 and internal consistency with Cronbach's α = >0.90 (22).

Validity. *Criterion validity.* A correlation of −0.72 has been shown between HAQ scores and physical capacity measures (23). Overall correlation with observed functional performance has been shown to be 0.88, with the lowest subcategory correlation 0.47 for arising, and the highest subcategory correlation 0.88 for walking (1). In patients with RA, the predominant determinants of HAQ disability are disease activity, pain, and psychosocial factors (24).

Construct validity. HAQ scores have been shown to correlate well with both clinical and laboratory measures, including joint counts and inflammatory markers (21,25). Construct validity has also been confirmed using cross-validation with exploratory factor analysis and confirmatory factor analysis (26,27).

Predictive validity. In RA, the HAQ is among the strongest predictors of long-term outcomes including work disability and economic loss (28,29). It has been shown to be the most important predictor of mortality, compared to other patient measures including radiographs, joint counts, and laboratory values (29).

Ability to detect change. Minimal clinically important differences for HAQ scores have been published at ~0.22, although estimates range widely (0.07–0.87) depending on the population and construct used (16,30–33). The HAQ has been shown to have high sensitivity (3 SD at a reliability >0.95), but is limited in the normal function range (8).

Critical Appraisal of Overall Value to the Rheumatology Community

For more than 30 years the HAQ has been the gold standard measure of functional disability in rheumatology. Comparisons with it are required to show validity in new measures, and all new treatments report change in HAQ to show efficacy. With RA being diagnosed earlier and as more patients have normal or better function, the HAQ floor effect (demonstrated by ~10% of RA patients who cannot improve in score despite clinical improvement) (3,5,34) has grown as an important limitation. Also, the relatively long length of the HAQ has led others to develop shorter, similar measures for clinical use.

MODIFIED HEALTH ASSESSMENT QUESTIONNAIRE (MHAQ)

Description

Purpose. The MHAQ was developed as a short version of the HAQ with the goal of decreased patient and provider time commitment (35). The MHAQ was developed for use in patients with rheumatic disease as an assessment of functional status. The MHAQ has also been used to assess function after joint arthroplasty (36).

Content. Eight items regarding daily activity, 1 from each of the 8 HAQ categories.

Number of items. There are 8 items, 1 from each of the 8 categories of the HAQ. The MHAQ does not address the use of aids or assistive devices.

Response options/scale. Eight activities are rated on a 4-point Likert scale where 0 = without any difficulty, 1 = with some difficulty, 2 = with much difficulty, and 3 = unable to do.

Recall period for items. Three months.

Practical Application

How to obtain. Developer contact information: Theodore Pincus, MD, New York University Hospital for Joint Diseases, 301 East 17th Street, New York, NY 10003. The MHAQ may be used free of charge from http://www.iche.edu/newsletter/MHAQ.pdf.

Method of administration. Most frequently self-administered, but may also be administered by an in-person or telephone interviewer.

Scoring. The MHAQ may be calculated by hand or with a calculator by adding all scored items together (at least 6 of the 8 items are required) and dividing by the total number of items answered to obtain the final score.

Score interpretation. Total score is between 0.0–3.0, in 0.125 increments. Higher scores indicate worse function and greater disability. MHAQ scores <0.3 are considered normal. It has been proposed that MHAQ scores be divided into categories of mild (MHAQ <1.3), moderate (1.3 < MHAQ < 1.8) and severe (MHAQ >1.8) functional losses. A change in MHAQ of 0.25 has been suggested as clinically meaningful (37). MHAQ scores are nonlinear and a change from 0 to 1 may not indicate the same amount of functional change as a change from 2 to 3 (38). A conversion formula exists to transform MHAQ into HAQ scores (39).

Respondent burden. Time to complete is <5 minutes.

Administrative burden. Time to administer is <5 minutes. Time to score is <1 minute. No specific training is necessary to score.

Translations/adaptations. Originally developed in English, the MHAQ has also been translated to selected additional languages (40,41).

Psychometric Information

Reliability. Repeated testing within 4–5 weeks has shown test–retest reliability for the MHAQ to be good–excellent (κ = 0.65–0.91, $P < 0.001$).

Validity. *Concurrent validity.* The MHAQ and HAQ have been shown to be highly correlated (r = 0.857), however, average MHAQ scores have been shown to be 0.58 lower than HAQ scores (39). Significant correlation between the responses of the individual MHAQ items and the included original HAQ items (r = 0.71–0.84, $P <$ 0.001), and between the remaining excluded HAQ items (r = 0.55–0.69, $P < 0.001$) have been shown (2). The MHAQ correlates with the physical function domain of the World Health Organization Quality of Life short version physical domain (r = −0.62, $P < 0.01$) (36).

Construct validity. Patients and providers may not agree on the importance of specific aspects of functional disability (κ = 0.16), however, patients and providers rank the importance of the MHAQ domains in the same order (43). While disability, as measured by the MHAQ, and patient satisfaction have been shown to correlate overall (r = 0.69, $P < 0.001$), individual patients may perceive the same level of disability with differing levels of satisfaction (2). Among the 8 items composing the MHAQ, no important differential item functioning has been identified. The MHAQ is a primarily unidimensional instrument measuring function, however, it is slightly 2-dimensional as it measures functional aspects of both upper and lower extremities (40).

Predictive validity. In one study, when combined with age and comorbidities, the MHAQ was shown to predict 5-year mortality better than radiographic and laboratory data (43). Composite and individual MHAQ item scores have been shown to be better correlated with changes in clinical variables (joint counts, grip strength, pain, morning stiffness, erythrocyte sedimentation rate, and patient global status) than HAQ change scores (44).

Ability to detect change. The MHAQ was found to be similarly sensitive to change with close correlation between change in the MHAQ and HAQ when administered monthly over time in a 12-month clinical trial comparing methotrexate and leflunomide therapy for rheumatoid arthritis (RA) (45). The MHAQ, however, lacks a normal distribution with up to 95% of values clustering between 0 and 1.5 (37) and may fail to detect numerical improvement in scores despite clinical improvement in up to 25% of patients (5). Addition of items has been shown to decrease this floor effect (41).

Critical Appraisal of Overall Value to the Rheumatology Community

The MHAQ was developed for use in the clinic as a shorter, more manageable version of the HAQ with simplified scoring methods. The MHAQ is often employed in outcomes research and clinical care as a substitute for the original HAQ and is a component of several composite RA disease activity measurement tools (i.e., Global Arthritis Score, Patient-Based Disease Activity Score). Unfortunately, the MHAQ is not equivalent to the original HAQ (39), with assessment of fewer items potentially missing the extent of functional impairment (35). MHAQ scores have been shown to lack sensitivity to change (34,46), are routinely lower than HAQ scores by 0.3–0.5 units (3,39,47), and tend to cluster at the lower end of the scale, leading to a non-normal distribution of values (5,34). The much larger floor effect may limit ~25% of all RA patients from having a change in the MHAQ even with clinical improvement (3,5,34,35,40). Another limitation of the MHAQ is that the assessment asks for the degree of change in difficulty with specific tasks over the preceding 3 months, and is therefore subject to recall bias, although it has been shown that the MHAQ is correlated with HAQ change scores. This same issue could however be considered an advantage of the MHAQ over the HAQ as repeated administration of the HAQ with calculation of change scores may be cumbersome (44).

MULTIDIMENSIONAL HEALTH ASSESSMENT QUESTIONNAIRE (MDHAQ)

Description

Purpose. The MDHAQ was originally developed as an assessment of functional status for use in patients with rheumatic disease. It is intended to be a shorter substitute for the HAQ with the goal of decreased patient and provider time commitment. The MDHAQ was designed to improve the ability to detect improvements in function at the lower end of the scale as compared to the MHAQ.

Content. Ten items regarding daily activity: the 8 MHAQ items plus "walk 2 miles" and "participate in recreational activities and sports as you would like."

Number of items. Ten. No subscales.

Response options/scale. Ten activities are rated on a 4-point Likert scale where 0 = without any difficulty, 1 = with some difficulty, 2 = with much difficulty, and 3 = unable to do.

Recall period for items. One week.

Practical Application

How to obtain. Developer contact information: Theodore Pincus, MD, New York University Hospital for Joint Diseases, 301 East 17th Street, New York, NY 10003. Available online at www.mdhaq.org.

Method of administration. Most frequently self-administered, but may also be administered by an in-person or telephone interviewer.

Scoring. The MDHAQ may be calculated by hand or with a calculator by adding all scored items together (at least 9 of the 10 items are required), dividing by the total number of items answered rounding to the nearest 0.1 to obtain a final score from 0–3. A calculator or an available scoring template may also be used to give a final score from 0–10 (48).

Score interpretation. Scores range from 0–3 and higher scores indicate worse function and greater disability.

Respondent burden. Time to complete is <5 minutes.

Administrative burden. Time to administer is <5 minutes. The MDHAQ takes <10 seconds to score (20). No specific training is necessary, however, scoring instructions for the MDHAQ may be found at www.mdhaq.org.

Translations/adaptations. Originally developed in English, the MDHAQ has been translated to selected additional languages (40,41,49–51).

Psychometric Information

Reliability. Good-to-excellent test–retest reliability with kappa scores for each item ranging from 0.65–0.81 (all $P < 0.001$) (3). Scores are replicable among young and geriatric populations, with more consistent reliability in subjects age <40 years compared to those age >65 years (Cronbach's $\alpha = 0.82$ versus 0.61) (52).

Validity. *Concurrent validity.* The MDHAQ and HAQ have been shown to be highly correlated; however, average MDHAQ scores have been shown to be 0.34 lower than HAQ scores (39).

Construct validity. MDHAQ scores correlate with the Disease Activity Score in 28 joints (DAS28) at baseline (r = 0.51), although the change in MDHAQ over 12 months correlated less well with change in the DAS28 (r = 0.39) (53). The 2 items regarding participation in sports and walking 2 miles do not fit Rasch model criteria for unidimensionality. In addition, missing MDHAQ items affect the final score more than for HAQ-II, but less than for HAQ or MHAQ (5).

Predictive validity. MDHAQ scores were more significantly associated with degree of morning stiffness than pain, fatigue, joint counts, and patient global (54). MDHAQ scores have also been shown to independently predict 10-year mortality among people with rheumatoid arthritis (55).

Ability to detect change. Variability of scores over time was not significantly different compared to variability of pain and patient global assessment scores ($P = 0.13$) in a study of weekly self-assessment over 6 months (56). Like other HAQ-variations, the MDHAQ lacks a normal distribution at the lower end of the scale and may fail to detect numerical improvement in scores despite clinical improvement in up to 4.4% of patients (5). Addition of items has been shown to decrease this floor effect (40).

Critical Appraisal of Overall Value to the Rheumatology Community

As a shorter version of the original HAQ, the MDHAQ was developed for use in the clinic to improve the ability of the MHAQ to detect functional improvement at the lower end of the scale. While only the 10-item functional scale of the MDHAQ is formally scored, the developers suggest administering the MDHAQ as a 2-page questionnaire (57) with inclusion of nonscored items relevant to patient care as the composite Rheumatoid Arthritis Prevention of Structural Damage (RAPID) scores, which measure rheumatologic disease activity (RAPID indices are covered elsewhere in this issue). As compared to the MHAQ, the MDHAQ is the same but with 2 difficult items added, which result in the MDHAQ having a lower chance of failure to detect numerical improvement when clinical improvement is present as compared with both the HAQ and MHAQ. Like other HAQ variants, the MDHAQ is scored between 0–3 for comparison with the original HAQ, and similarly to the HAQ it deviates from a normal distribution at values near zero (3,57). Additionally, the MDHAQ has more even spacing of items than the HAQ and MHAQ, making a change of 0.5 more similar across the range of the scale (57), although outliers remain and item spacing is inferior to that of the HAQ-II (5).

HEALTH ASSESSMENT QUESTIONNAIRE II (HAQ-II)

Description

Purpose. The HAQ-II was developed to assess functional status in individuals with rheumatic disease. It is intended to be a short replacement for the HAQ and was created using an item bank and Rasch analysis to best balance item fit, scale length, and item spacing in an attempt to correct the floor effects seen with earlier modifications of the HAQ (5).

Content. Ten items regarding daily activity: toileting, opening doors, standing from a chair, walking on flat ground, waiting in line, reaching for an object, ambulating up steps, performing outdoor work, lifting heavy objects, and moving heavy objects.

Number of items. Ten items are included, 5 from the original HAQ and 5 additional items, all in the form of questions addressing functional ability. No subscales.

Response options/scale. Ten activities are rated on a 4-point Likert scale where 0 = without any difficulty, 1 = with some difficulty, 2 = with much difficulty, and 3 = unable to do.

Recall period for items. One week.

Practical Application

How to obtain. Developer contact information: Frederick Wolfe, MD, National Data Bank for Rheumatic Diseases, 1035 North Emporia Avenue, Suite 288, Wichita, KS 67214. The HAQ-II may be used free of charge and is available at http://www.arthritis-research.org/research/HAQ-II.

Method of administration. Most frequently self-administered, but may also be administered by an in-person or telephone interviewer.

Scoring. The HAQ-II may be calculated by hand or with a calculator by adding all scored items together (at least 8 of the 10 items are required) and dividing by the total number of items answered to obtain the final score.

Score interpretation. Total score can range from 0–3.0, in 0.1 increments and higher scores indicate worse function and greater disability. HAQ-II scores are nonlinear and a change from 0 to 1 may not indicate the same amount of functional change as a change from 2 to 3 (38). Conversion formula exist to transform HAQ-II into HAQ scores (39).

Respondent burden. Time to complete is <5 minutes.

Administrative burden. Time to administer is <5 minutes. Time to score is <1 minute. No specific training is necessary.

Translations/adaptations. Originally developed in English, the HAQ-II has been translated to Dutch (58).

Psychometric Information

Reliability. The HAQ-II demonstrates satisfactory reliability (Cronbach's α = 0.88) (5). Test–retest reliability studies have not been performed.

Validity. *Concurrent validity.* The HAQ-II and HAQ are highly correlated (r = 0.92) (5) with average HAQ-II scores shown to be only minimally lower (by 0.02–0.04) than HAQ scores (39). Notably, HAQ and HAQ-II have been shown not to be interchangeable in an individual patient (5). At similar levels to the HAQ, the HAQ-II was shown to correlate with the Short Form 36 physical function scale (r = −0.85) and EuroQol utility scales (r = −0.67), and correlate positively with the Rheumatology Distress Index (r = 0.61), the Rheumatoid Arthritis Disease Activity Index (r = 0.65), the Work Limitations Questionnaire Index (r = 0.56), and the Arthritis Impact Measurement Scales depression and anxiety scales (r = 0.44 and 0.38, respectively) (5).

Construct validity. The HAQ-II was designed using Rasch analysis and was found to measure disability over a longer scale than the HAQ, and has no nonfitting items and no gaps between items (5). The HAQ-II is not a true unidimensional tool and includes 9 items assessing functional limitations and 1 measure of disability ("doing outside work") (5).

Predictive validity. HAQ-II values are correlated with clinical outcomes including pain, fatigue, patient's and physician's assessments of global disease severity, Disease Activity Score in 28 joints, erythrocyte sedimentation rate, joint counts, medical costs, joint replacement, and work disability, at levels similar to those of the HAQ and MHAQ (5).

Ability to detect change. Like the HAQ, the HAQ-II lacks a normal distribution at values near 0 and may fail to detect numerical improvement in scores despite clinical improvement in up to 5.8% of patients (5).

Critical Appraisal of Overall Value to the Rheumatology Community

The HAQ-II is a 10-item functional questionnaire based on the original HAQ, with scores ranging from 0–3. It is easily administered in the clinical setting and is suitable for use in studies when a HAQ substitute is required. Of the HAQ versions, the HAQ-II has been shown to have the greatest uniformity between values across the range of the scale and provides the least impact on total score for skipped items (5,58). While shorter and simpler than the HAQ, the HAQ-II has demonstrated similar levels of reliability and validity, is more closely correlated with the original HAQ than other HAQ modifications (39), and requires the least manipulation of data in order to compare with the original HAQ (39). Furthermore, the HAQ-II has a lesser floor effect, as compared with the HAQ and MHAQ, with potential failure to detect clinical improvement in only 5.8% of patients (5,31,59) as compared to 10% for the HAQ (3,5,34) and up to 25% for the MHAQ (3,5,34,35,40).

IMPROVED HAQ

Description

Purpose. To measure current level of difficulty in performing activities of daily living. A slightly modified version of the HAQ, this measure uses the same contextual 20 items to assess activities of daily living, but in the present tense and adds "with a little bit of difficulty" as an additional response option to reduce floor effects. First introduced in 2007 as the HAQ-100 and then the PROMIS HAQ, its name has been revised by its developers to the Improved HAQ in an effort to avoid confusion with official instruments of the Patient-Reported Outcomes Measurement Information System (PROMIS) component of the National Institutes of Health Roadmap Initiative. Developed in both general population, rheumatoid arthritis and osteoarthritis patients, the Improved HAQ remained targeted for those with some physical limitations (8). The developers recommend the use of the Improved HAQ in clinical research in all instances where the HAQ would have been used, except for studies in progress where baseline measurements used the original HAQ (4).

Content. Twenty questions assessing current ability to perform specific activities of daily living. Four questions assessing use of assistive devices or help from another person (aids/help) in the categories of walking, standing, dressing, and reaching.

Number of items. There are 24 total items: 20 items covering activities of daily living and 4 questions regarding use of aids/help. In contrast to the original HAQ, items are not grouped by physical function category.

Response options/scale. Twenty items are asked on a 5-point Likert scale where 0 = without any difficulty, 1 = with a little difficulty, 2 = with some difficulty, and 3 = with much difficulty, and 4 = unable to do. Four additional yes/no questions ask about specific use of aids/help.

Recall period for items. One week.

Practical Application

How to obtain. Developer contact information: James F. Fries, MD, Division of Immunology and Rheumatology, Stanford University Medical Center, 1000 Welch Road, Suite 203, Palo Alto, CA 94304-1808. Available at http://aramis.stanford.edu/HAQ.html.

Method of administration. Most frequently self-administered, but may also be administered by an in-person or telephone interviewer.

Scoring. By hand or electronically by first adjusting item scores based on any use of aids/help by increasing those items to a minimum of 3 (out of 4) using a provided table. Add all scored items together (at least 15 of the 20 items are required) and divide by the total number of items answered to obtain a score between 0 and 4. Finally, multiply this score by 25 and round to the nearest whole number. Alternative scoring methods are being evaluated (4).

Score interpretation. Score range of 0–100. Higher scores reflect greater functional limitation.

Respondent burden. Time to complete is ~5 minutes.

Administrative burden. Time to administer is ~5 minutes. Time to score is ~2 minutes. Familiarity with the adjustment table for use of aids/help improves scoring efficiency.

Translations/adaptations. A 16-item Improved HAQ has been proposed (4).

Psychometric Information

Reliability. Is slightly better than the HAQ with a reliability >0.95 for most of the range of decreased physical function covered (8).

Validity. For concurrent validity, correlations between component items vary from 0.37–0.75 ($P < 0.0001$), and scores were very similar to the HAQ when compared on the same scale (mean ± SD Improved HAQ mean 25 ± 25 versus mean ± SD HAQ 24 ± 23) (8).

Ability to detect change. In a large population not limited to arthritis, the Improved HAQ has reduced floor effects as compared with the HAQ (% patients scoring 0: 66.3% versus 73.2%; $P < 0.05$) (8).

Critical Appraisal of Overall Value to the Rheumatology Community

Developed as an updated HAQ with improved psychometric properties through use of item response theory and qualitative analysis, the Improved HAQ only slightly differs from the original HAQ. Twenty items remain, although the original 21 aids/help questions have been reduced to 4. The relatively small quality improvement gained with the Improved HAQ while retaining all 20 items may make it slow to be adopted in the clinic or in research. The change from a 0–3 to a 0–100 scale also makes comparison with prior HAQ studies nonintuitive, and there is no formula yet to transform scores to the HAQ for research or continued clinical care. Although there have been limited published studies using the new Improved HAQ, it is easy to assume it would have similar or better construct and predictive validity of the HAQ. It is currently unclear how much the Improved HAQ will be used in place of the HAQ with its extensive experience as the gold standard, but it provides a useful tool in further development of functional measures.

RHEUMATOID ARTHRITIS QUALITY OF LIFE (RAQoL)

Description

Purpose. The RAQoL is a disease-specific measure that assesses self-reported quality of life in patients with rheumatoid arthritis (RA) (60,61).

Content. Questions assess specific activities of daily living and quality of life (6).

Number of items. 30.

Response options/scale. Each item is answered with 1) yes or 2) no.

Recall period for items. One week.

Practical Application

How to obtain. Developer contact information: Diane Whalley, Galen Research, Enterprise House, Manchester Science Park, Lloyd Street North, Manchester M15 6SE, United Kingdom (6). Available in De Jong et al (6).

Method of administration. Self-assessment using a questionnaire format is preferred by the developer to avoid introducing an additional source for experimental error (6), and it may also be administered by an in-person or telephone interviewer.

Scoring. The number of items answered "yes" are totaled, giving the final score.

Score interpretation. Score range of 0–30. Higher scores indicate worsening quality of life.

Respondent burden. Time to complete is 2–8 minutes (6).

Administrative burden. Time to administer is 2–8 minutes (6). Time to score is <1 minute. No specific training is necessary.

Translations/adaptations. Originally developed in the UK and The Netherlands, the RAQoL has additionally been developed for use in Turkish, Canadian, Estonian, Australian, and Swedish RA populations (6,61–66).

Psychometric Information

Reliability. Test–retest reliability was shown to be excellent with Spearman's rank correlation coefficient >0.90 (6). After repeated testing at 2 and 12 weeks apart, the intraclass correlation coefficient was 0.79 (22) and 0.99 (67), respectively. Internal consistency was also shown to be excellent with Cronbach's $\alpha = 0.92$–0.94 (6,22,67).

Validity. *Concurrent validity.* RAQoL scores correlate with other measures of quality of life in RA cohorts including a rating scale utility (r = −0.63) and EuroQol 5-domain (EQ-5D) (r = −0.62 to −0.76) (6,22,68,69).

Content validity. Correlation with domains of the Nottingham Health Profile (NHP) shows strong relationships between RAQoL and physical mobility, energy level, and pain (6). The RAQoL also demonstrates correlations with the Disease Activity Score (r = 0.41–0.82), pain (r = 0.48–0.86), Health Assessment Questionnaire (HAQ) (r = 0.73–0.86), patient global (r = 0.62–0.82), fatigue (r = 0.78), swollen and tender joint count (r = 0.53), modified Sharp score (r = 0.38) and physician global (r = 0.36) (16,22,68–70).

Construct validity. Principal component analysis revealed 4 primary factors with high reliability (Cronbach's alpha): mobility/energy 0.79, self-care 0.75, mood/emotion 0.71, and physical contact 0.54 (69). There was a medium effect size (ES; 0.71) for distinguishing those on disability pension (22) and a large ES (0.81) for patients that took days off of work in the previous year due to RA (16).

Ability to detect change. Responsiveness measured by standardized response mean was shown to be −0.67 to −0.51 for patients reporting improvement over 6 months, −0.16 for no change, and 0.18 for deterioration (21,71). Minimally important worsening of physical function as measured by the HAQ (0.25 increase on HAQ) corresponds to an increase of 2.0 in RAQoL score (16,72). A statistically significant response was found 12 weeks after 37 patients initiated biologic therapy (ES −1.13) (67); similarly, a 29% improvement (14–10 score) was shown 12 months after 126 patients initiated biologic therapy (73).

Critical Appraisal of Overall Value to the Rheumatology Community

The RAQoL has consistently shown good responsiveness and validity as a quality of life measure in RA patients. While quality of life measurement in RA is useful and needed to help justify the high costs of new RA therapies, there has been relatively little use of the RAQoL in clinical trials or drug efficacy studies. Most cost-effectiveness studies in RA have relied upon mapping changes in HAQ on indirect health utilities, and there have not yet been any attempts to map the RAQoL similarly. The "physical contact" dimension is unique to the RAQoL and represents an important RA patient concern of avoiding shaking hands or being touched. The greatest limitation of the RAQoL may be its length, i.e., 30 items for an RA-specific measure is hard to justify when there are popular and psychometrically similar generic utility measures with 5 (EQ-5D) to 36 (Short Form 36 Health Survey) items.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Abbott Laboratories had no financial interest in this project and had no input in the design or content, and all opinions and conclusions expressed herein are those of the authors.

REFERENCES

1. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
2. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
3. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum 1999;42:2220–30.
4. Fries JF, et al. The Arthritis, Rheumatism, and Aging Medical Information System. Aramis: HAQ. URL: http://aramis.stanford.edu/HAQ.html.
5. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum 2004;50:3296–305.
6. De Jong Z, van der Heijde D, McKenna SP, Whalley D. The reliability and construct validity of the RAQoL: a rheumatoid arthritis-specific quality of life instrument. Br J Rheumatol 1997;36:878–83.
7. Wolfe F, Michaud K, Pincus T. Preliminary evaluation of a visual analog function scale for use in rheumatoid arthritis. J Rheumatol 2005;32:1261–6.
8. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061–6.
9. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: dimensions and practical applications. Health Qual Life Outcomes 2003;1:20.
10. Lubeck DP, Fries JF. Health status among persons infected with human immunodeficiency virus: a community-based study. Med Care 1993;31:269–76.
11. Gillen M. Injuries from construction falls: functional limitations and return to work. AAOHN J 1999;47:65–73.
12. Carreno A, Fernandez I, Badia X, Varela C, Roset M. Using HAQ-DI to estimate HUI-3 and EQ-5D utility values for patients with rheumatoid arthritis in Spain. Value Health 2011;14:192–200.
13. Wolfe F, Michaud K, Wallenstein G. Scale characteristics and mapping accuracy of the US EQ-5D, UK EQ-5D, and SF-6D in patients with rheumatoid arthritis. J Rheumatol 2010;37:1615–25.
14. Standfield L, Norris S, Harvey C, Elliot L, Riordan J, Hall S, et al. Relationship between rheumatoid arthritis disease severity, health-related utility, and resource use in Australian patients: a cross-sectional, multicenter study. Clin Ther 2010;32:1329–42.
15. Bansback N, Marra C, Tsuchiya A, Anis A, Guh D, Hammond T, et al. Using the health assessment questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis. Arthritis Rheum 2007;57:963–71.
16. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60:1571–82.
17. Kobelt G, Lindgren P, Lindroth Y, Jacobson L, Eberhardt K. Modelling the effect of function and disease activity on costs and quality of life in rheumatoid arthritis. Rheumatology (Oxford) 2005;44:1169–75.
18. Kobelt G, Jonsson L, Young A, Eberhardt K. The cost-effectiveness of infliximab (Remicade) in the treatment of rheumatoid arthritis in Sweden and the United Kingdom based on the ATTRACT study. Rheumatology (Oxford) 2003;42:326–35.

19. Tomlin GS, Holm MB, Rogers JC, Kwoh CK. Comparison of standard and alternative health assessment questionnaire scoring procedures for documenting functional outcomes in patients with rheumatoid arthritis. J Rheumatol 1996;23:1524–30.
20. Yazici Y, Bergman M, Pincus T. Time to score quantitative rheumatoid arthritis measures: 28-Joint Count, Disease Activity Score, Health Assessment Questionnaire (HAQ), Multidimensional HAQ (MDHAQ), and Routine Assessment of Patient Index Data (RAPID) scores. J Rheumatol 2008;35:603–9.
21. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 2nd ed. New York: Oxford University Press; 1996. p. 106–15.
22. Linde L, Sorensen J, Ostergaard M, Horslev-Petersen K, Hetland ML. Health-related quality of life: validity, reliability, and responsiveness of SF-36, 15D, EQ-5D (corrected) RAQoL, and HAQ in patients with rheumatoid arthritis. J Rheumatol 2008;35:1528–37.
23. Daltroy LH, Larson MG, Eaton HM, Phillips CB, Liang MH. Discrepancies between self-reported and observed physical function in the elderly: the influence of response shift and other factors. Soc Sci Med 1999;48:1549–61.
24. Wolfe F. A reappraisal of HAQ disability in rheumatoid arthritis. Arthritis Rheum 2000;43:2751–61.
25. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.
26. Sousa KH, Kwok OM, Ryu E, Cook SW. Confirmation of the validity of the HAQ-DI in two populations living with chronic illnesses. J Nurs Meas 2008;16:31–42.
27. Cole JC, Motivala SJ, Khanna D, Lee JY, Paulus HE, Irwin MR. Validation of single-factor structure and scoring protocol for the Health Assessment Questionnaire-Disability Index. Arthritis Rheum 2005;53: 536–42.
28. Wolfe F, Hawley DJ. The longterm outcomes of rheumatoid arthritis: work disability: a prospective 18 year study of 823 patients. J Rheumatol 1998;25:2108–17.
29. Wolfe F, Michaud K, Gefeller O, Choi HK. Predicting mortality in patients with rheumatoid arthritis. Arthritis Rheum 2003;48:1530–42.
30. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557–60.
31. Pope JE, Khanna D, Norrie D, Ouimet JM. The minimally important difference for the health assessment questionnaire in rheumatoid arthritis clinical practice is smaller than in randomized controlled trials. J Rheumatol 2009;36:254–9.
32. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE Jr. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum 2000;43:1478–87.
33. Wolfe F, Michaud K, Strand V. Expanding the definition of clinical differences: from minimally clinically important differences to really important differences: analyses in 8,931 patients with rheumatoid arthritis. J Rheumatol 2005;32:583–9.
34. Stucki G, Stucki S, Bruhlmann P, Michel BA. Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. Ann Rheum Dis 1995;54: 461–5.
35. Wolfe F. Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. J Rheumatol 2001;28:982–9.
36. Ackerman IN, Graves SE, Bennell KL, Osborne RH. Evaluating quality of life in hip and knee replacement: psychometric properties of the World Health Organization Quality of Life short version instrument. Arthritis Rheum 2006;55:583–90.
37. Wolfe F, Pincus T. Listening to the patient: a practical guide to self-report questionnaires in clinical care. Arthritis Rheum 1999;42: 1797–808.
38. Wolfe F. The psychometrics of functional status questionnaires: room for improvement. J Rheumatol 2002;29:865–8.
39. Anderson J, Sayles H, Curtis JR, Wolfe F, Michaud K. Converting modified health assessment questionnaire (HAQ), multidimensional HAQ, and HAQII scores into original HAQ scores using models developed with a large cohort of rheumatoid arthritis patients. Arthritis Care Res (Hoboken) 2010;62:1481–8.
40. Martin M, Kosinski M, Bjorner JB, Ware JE Jr, Maclean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res 2007;16:647–60.
41. Calvo-Alen J, Corrales A, Sanchez-Andrada S, Fernandez-Echevarria MA, Pena JL, Rodriguez-Valverde V. Functional outcome and subset identification in RA patients from meridional Europe: analysis of a Spanish cohort. Clin Rheumatol 2003;22:77–83.
42. Hewlett S, Smith AP, Kirwan JR. Values for function in rheumatoid arthritis: patients, professionals, and public. Ann Rheum Dis 2001;60: 928–33.
43. Callahan LF, Pincus T, Huston JW 3rd, Brooks RH, Nance EP Jr, Kaye JJ. Measures of activity and damage in rheumatoid arthritis: depiction of changes and prediction of mortality over five years. Arthritis Care Res 1997;10:381–94.
44. Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ. Ann Rheum Dis 1992;51:1202–5.
45. Tugwell P, Wells G, Strand V, Maetzel A, Bombardier C, Crawford B, et al, on behalf of the Leflunomide Rheumatoid Arthritis Investigators Group. Clinical improvement as reflected in measures of function and health-related quality of life following treatment with leflunomide compared with methotrexate in patients with rheumatoid arthritis: sensitivity and relative efficiency to detect a treatment effect in a twelve-month, placebo-controlled trial. Arthritis Rheum 2000;43: 506–14.
46. Serrano MA, Beltran Fabregat J, Olmedo Garzon J. Should the MHAQ ever be used? [letter]. Ann Rheum Dis 1996;55:271–2.
47. Uhlig T, Kvien TK, Glennas A, Smedstad LM, Forre O. The incidence and severity of rheumatoid arthritis: results from a county register in Oslo, Norway. J Rheumatol 1998;25:1078–84.
48. Pincus T, Yazici Y, Bergman M. A practical guide to scoring a Multi-Dimensional Health Assessment Questionnaire (MDHAQ) and Routine Assessment of Patient Index Data (RAPID) scores in 10-20 seconds for use in standard clinical care, without rulers, calculators, websites or computers. Best Pract Res Clin Rheumatol 2007;21:755–87.
49. Lee SS, Park MJ, Yoon HJ, Park YW, Park IH, Park KS. Evaluating the Korean version of the Multidimensional Health Assessment Questionnaire in patients with rheumatoid arthritis. Clin Rheumatol 2006;25: 353–7.
50. Arkela-Kautiainen M, Kautiainen H, Uutela T, Laiho K, Blafield H, Leirisalo-Repo M, et al. Evaluation of the MDHAQ in Finnish patients with RA (corrected). J Rheumatol 2005;32:1426–31.
51. Kumar A, Malaviya AN, Pandhi A, Singh R. Validation of an Indian version of the Health Assessment Questionnaire in patients with rheumatoid arthritis. Rheumatology (Oxford) 2002;41:1457–9.
52. Doventas A, Karadag B, Curgunlu A, Bilici A, Sut N, Erdincler DS, et al. Replicability and reliability of pain assessment forms in geriatrics. Arch Gerontol Geriatr 2010. E-pub ahead of print.
53. Sullivan MB, Iannaccone C, Cui J, Lu B, Batra K, Weinblatt M, et al. Evaluation of selected rheumatoid arthritis activity scores for office-based assessment. J Rheumatol 2010;37:2466–8.
54. Yazici Y, Pincus T, Kautiainen H, Sokka T. Morning stiffness in patients with early rheumatoid arthritis is associated more strongly with functional disability than with joint swelling and erythrocyte sedimentation rate. J Rheumatol 2004;31:1723–6.
55. Pincus T, Keysor J, Sokka T, Krishnan E, Callahan LF. Patient questionnaires and formal education level as prospective predictors of mortality over 10 years in 97% of 1,416 patients with rheumatoid arthritis from 15 United States private practices. J Rheumatol 2004;31: 229–34.
56. Blanchais A, Berthelot JM, Fontenoy AM, le Goff B, Maugars Y. Weekly home self-assessment of RAPID-4/3 scores in rheumatoid arthritis: a 6-month study in 26 patients. Joint Bone Spine 2010;77:582–7.
57. Pincus T, Sokka T, Kautiainen H. Further development of a physical function scale on a MDHAQ (corrected) for standard care of patients with rheumatic diseases. J Rheumatol 2005;32:1432–9.
58. Ten Klooster PM, Taal E, van de Laar MA. Rasch analysis of the Dutch health assessment questionnaire disability index and the health assessment questionnaire II in patients with rheumatoid arthritis. Arthritis Rheum 2008;59:1721–8.
59. Wolfe F. Why the HAQ-II can be an effective substitute for the HAQ. Clin Exp Rheumatol 2005;23 Suppl 39:S29–30.
60. Lillegraven S, Kvien TK. Measuring disability and quality of life in established rheumatoid arthritis. Best Pract Res Clin Rheumatol 2007; 21:827–40.
61. Whalley D, McKenna SP, de Jong Z, van der Heijde D. Quality of life in rheumatoid arthritis. Br J Rheumatol 1997;36:884–8.
62. Kutlay S, Kucukdeveci AA, Gonul D, Tennant A. Adaptation and validation of the Turkish version of the Rheumatoid Arthritis Quality of Life Scale. Rheumatol Int 2003;23:21–6.
63. Neville C, Whalley D, McKenna S, Le Comte M, Fortin PR. Adaptation and validation of the rheumatoid arthritis quality of life scale for use in Canada. J Rheumatol 2001;28:1505–10.
64. Eberhardt K, Duckberg S, Larsson BM, Johnson PM, Nived K. Measuring health related quality of life in patients with rheumatoid arthritis: reliability, validity, and responsiveness of a Swedish version of RAQoL. Scand J Rheumatol 2002;31:6–12.
65. Tammaru M, McKenna SP, Meads DM, Maimets K, Hansen E. Adaptation of the rheumatoid arthritis quality of life scale for Estonia. Rheumatol Int 2006;26:655–62.

66. Cox SR, McWilliams L, Massy-Westropp N, Meads DM, McKenna SP, Proudman S. Adaptation of the RAQoL for use in Australia. Rheumatol Int 2007;27:661–6.
67. Greenwood MC, Hakim AJ, Doyle DV. A simple extension to the Rheumatoid Arthritis Quality of Life Questionnaire (RAQoL) to explore individual patient concerns and monitor group outcome in clinical practice. Rheumatology (Oxford) 2006;45:61–5.
68. Inotai A, Rojkovich B, Fulop A, Jaszay E, Agh T, Meszaros A. Health-related quality of life and utility in patients receiving biological and non-biological treatments in rheumatoid arthritis. Rheumatol Int 2011. E-pub ahead of print.
69. Tijhuis GJ, de Jong Z, Zwinderman AH, Zuijderduin WM, Jansen LM, Hazes JM, et al. The validity of the Rheumatoid Arthritis Quality of Life (RAQoL) questionnaire. Rheumatology (Oxford) 2001;40:1112–9.
70. Garip Y, Eser F, Bodur H. Health-related quality of life in rheumatoid arthritis: comparison of RAQoL with other scales in terms of disease activity, severity of pain, and functional status. Rheumatol Int 2010. E-pub ahead of print.
71. Marra CA, Rashidi AA, Guh D, Kopec JA, Abrahamowicz M, Esdaile JM, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res 2005;14:1333–44.
72. Pentek M, Szekanecz Z, Czirjak L, Poor G, Rojkovich B, Polgar A, et al. Impact of disease progression on health status, quality of life and costs in rheumatoid arthritis in Hungary. Orv Hetil 2008;149:733–41. In Hungarian.
73. Raterman HG, Hoving JL, Nurmohamed MT, Herenius MM, Sluiter JK, Lems WF, et al. Work ability: a new outcome measure in rheumatoid arthritis? Scand J Rheumatol 2010;39:127–31.

Summary Table of Measures of Physical Function in Rheumatoid Arthritis*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence*	Validity evidence†	Ability to detect change†	Strengths	Cautions
HAQ	Assess limitations in physical function over past week	Patient self-report or interview	~5 minutes 41 questions	<2 minutes	Range 0–3; higher scores reflect worse function	Excellent	Excellent	Good	Gold standard for trial, clinical and research use, many translations	Lengthy, complex scoring, floor effect
MHAQ	Assess limitations in physical function over past 3 months	Patient self-report or interview	<5 minutes 8 questions	≤1 minute	Range 0–3; higher scores reflect worse function	Good	Good	Poor	Shortest in length, simplified scoring	Largest floor effect, group mean ~0.6 < HAQ, potential for recall bias
MDHAQ	Assess limitations in physical function over past week	Patient self-report or interview	<5 minutes 10 questions	≤1 minute	Range 0–3; higher scores reflect worse function	Good	Excellent	Fair	Short, simple scoring; more even distribution of scores than MHAQ	Floor effect, group mean ~0.3 < HAQ, unequal question difficulty
HAQ-II	Assess limitations in physical function over past week	Patient self-report or interview	<5 minutes 10 questions	≤1 minute	Range 0–3; higher scores reflect worse function	Good	Excellent	Good	Short Most even spacing of questions among HAQ versions	Group mean ~0.03 < HAQ, floor effect
Improved HAQ	Assess immediate limitations in physical function	Patient self-report or interview	~5 minutes 20 questions	≤2 minutes	Range 0–3; higher scores reflect worse function	Excellent	None available	Good	Improved psychometrics compared to HAQ	Lengthy, complex scoring, difficult to compare due to 0–100 range
RAQoL	Assess quality of life in RA	Patient self-report or interview	~6 minutes 30 questions	≤1 minute	Range 0–30; higher scores reflect poorer quality of life	Excellent	Good	Fair	Best validated RA-specific QoL measure Includes physical contact domain	Lengthy, lacks mapping to health utility

* HAQ = Health Assessment Questionnaire; MHAQ = Modified HAQ; MDHAQ = Multidimensional HAQ; RAQoL = Rheumatoid Arthritis Quality of Life.
† Subjective ratings (excellent, good, fair, poor) of psychometric properties were made by the authors based on evidence detailed in the text.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S64–S85
DOI 10.1002/acr.20577

Measures of Psoriatic Arthritis

Tender and Swollen Joint Assessment, Psoriasis Area and Severity Index (PASI), Nail Psoriasis Severity Index (NAPSI), Modified Nail Psoriasis Severity Index (mNAPSI), Mander/Newcastle Enthesitis Index (MEI), Leeds Enthesitis Index (LEI), Spondyloarthritis Research Consortium of Canada (SPARCC), Maastricht Ankylosing Spondylitis Enthesis Score (MASES), Leeds Dactylitis Index (LDI), Patient Global for Psoriatic Arthritis, Dermatology Life Quality Index (DLQI), Psoriatic Arthritis Quality of Life (PsAQOL), Functional Assessment of Chronic Illness Therapy–Fatigue (FACIT-F), Psoriatic Arthritis Response Criteria (PsARC), Psoriatic Arthritis Joint Activity Index (PsAJAI), Disease Activity in Psoriatic Arthritis (DAPSA), and Composite Psoriatic Disease Activity Index (CPDAI)

PHILIP J. MEASE

INTRODUCTION

The approaches to assessment of psoriatic arthritis (PsA) have matured significantly over the last decade due to the need for reliable measures in clinical trials. Additionally, there is a growing interest in a "treat to target" paradigm in the management of rheumatic diseases, i.e., the goal of achieving minimal disease activity or remission in order to maximize clinical improvement and minimize long-term damage, which requires quantitation of disease activity through validated measures. This paradigm has gained interest because of the increased understanding that has come from trials and clinical registries of patients with rheumatoid arthritis (RA) about the value of inhibiting the impact of disease symptoms and structural damage on function, quality of life, and long-term adverse outcomes related to comorbidities such as cardiovascular disease (1–4). It is becoming apparent that a similar value of tight control and treating to target exists in the management of PsA (5,6).

Work on outcome measures has been accomplished both in individual centers and the collaborative efforts of these centers through the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) and the Outcome Measures in Rheumatology (OMERACT) associations.

The principle clinical features of PsA to be assessed on physical examination include joint, skin and nail, entheseal, and spine disease, as well as dactylitis, which results from synovitis, tenosynovitis, and enthesitis. Disease activity in these domains may wax and wane in concert, or be divergent, contributing to the somewhat greater complexity of assessment of PsA as compared to RA. Efforts are underway to determine if composite measures of disease activity and response to therapy can be developed that effectively encompass all of these domains. Key domains assessed by patient-reported outcomes include pain, patient global, function, quality of life, and fatigue. Many of the measures of these clinical domains have been successfully adapted from measures used in the assessment of RA, ankylosing spondylitis (AS), and psoriasis. As such, many of these measures will only be touched on briefly in this article, with a focus on their adaptation to PsA, and the reader will be referred to their more extensive description of the basic measure in other articles in this issue. Measures that are not addressed elsewhere in the issue will be described in more detail in this article.

Through analysis of randomized controlled clinical trials, registry data sets, and expert Delphi exercises, a GRAPPA–OMERACT recommendation for a core set of domains to be assessed in PsA clinical trials has been established (7) (Figure 1). The set that should be assessed in all clinical trials (inner ring, Figure 1) includes peripheral joint activity, skin activity, patient global, pain, phys-

Philip J. Mease, MD: Seattle Rheumatology Associates, Swedish Medical Center, and University of Washington School of Medicine, Seattle.

Dr. Mease has received consultant fees, speaking fees, and/or honoraria (less than $10,000 each) from Centocor, BiogenIdec, Roche, UCB, Celgene, and Novartis, and (more than $10,000 each) from Abbott, Amgen, BMS, Lilly, and Pfizer.

Address correspondence to Philip J. Mease, MD, 1101 Madison Street, Suite 1000, Seattle, WA 98104. E-mail: pmease@nwlink.com.

Submitted for publication April 28, 2011; accepted in revised form July 22, 2011.

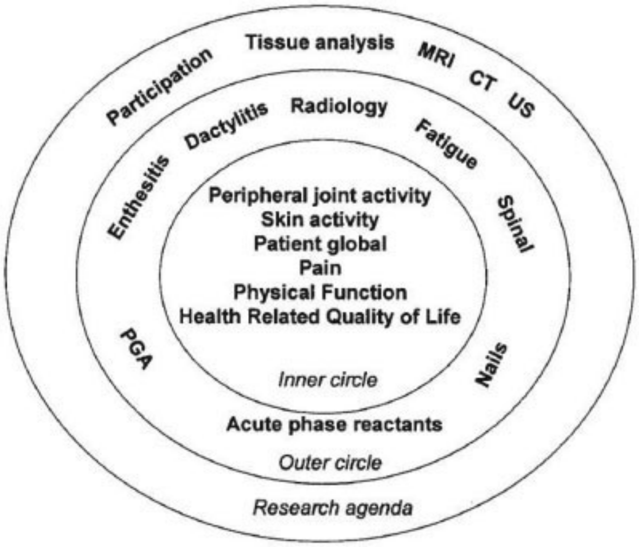


Figure 1. Domains for psoriatic arthritis (7). PGA = physician global assessment; MRI = magnetic resonance imaging; CT = computed tomography; US = ultrasound.

ical function, and health-related quality of life. Additional domains that ideally should be assessed at some point in a clinical development program include enthesitis, dactylitis, spine disease, nail disease, fatigue, physician global, acute-phase reactants, and structural status by radiograph (second ring, Figure 1). Additional measures still considered in the "research" area of assessments in development (outer ring, Figure 1) are computed tomography, magnetic resonance imaging, and ultrasound imaging scoring systems; tissue analysis (e.g., skin and synovial biopsy); and "participation" (ability to participate in meaningful life activities). The measures used to assess the GRAPPA–OMERACT domain set will be sequentially described in the remainder of this article. References to trials in which the measures have been used are summarized in recent reviews of PsA treatment (5,8,9). Many of these measures have not been officially validated in PsA, nor specifics of performance characteristics evaluated, other than the demonstration of their responsiveness and discriminant ability in clinical trials; therefore, psychometric characterization is not available for review for many measures.

The assessment of PsA has been aided by the development and utilization of formal classification criteria, which are used to select appropriate patients for clinical trials and registries. The criteria of the Classification of Psoriatic Arthritis Study Group were developed from an in-depth clinical, laboratory, and radiographic study of 588 PsA cases and 536 controls with RA, AS, or undifferentiated arthritis (10), using methods of logistic regression analysis, latent class analysis, classification and regression trees methodology, and receiver operating characteristic curve analysis. A patient qualifies for the criteria if they display inflammatory arthritis, enthesitis, and/or spondylitis and 3 points from a list of associated elements (Table 1). The criteria yielded a specificity of 98.7% and sensitivity of 91.4%, superior specificity than previously developed criteria such as those of Moll and Wright or Vasey and Espinoza (10).

GRAPPA is currently initiating an exercise to derive simple clinical definitions for what constitutes "inflammatory" arthritis, enthesitis, and spondylitis in order to aid nonrheumatologists as they attempt to distinguish inflammatory from noninflammatory forms of these conditions (Mease PJ: unpublished observations).

PERIPHERAL JOINT ASSESSMENT

A hallmark feature of psoriatic arthritis (PsA) is the presence of inflammatory arthritis, characterized by tenderness and or swelling due to synovial inflammation. Unlike rheumatoid arthritis (RA), wherein symmetric and polyarticular involvement is frequently seen, PsA may present in an oligoarticular and sometime monarticular pattern, often asymmetric, with a tendency to gradually become more

Table 1. CASPAR criteria (10)*

1. Psoriasis	
a. Current	Psoriatic skin or scalp disease present today as judged by a rheumatologist or dermatologist†
b. History	A history of psoriasis that may be obtained from patient, family doctor, dermatologist, or rheumatologist
c. Family history	A history of psoriasis in a first- or second-degree relative according to patient report
2. Psoriatic nail involvement	Typical psoriatic nail dystrophy, including onycholysis, pitting, and hyperkeratosis, observed on current physical examination
3. A negative test for RF	By any method except latex, but preferably by enzyme-linked immunosorbent assay or nephelometry, according to the local laboratory reference range
4. Dactylitis	
a. Current	Swelling of an entire finger
b. History	A history of dactylitis recorded by a rheumatologist
5. Radiologic evidence of juxtaarticular new bone formation	Ill-defined ossification near joint margins (but excluding osteophyte formation) on plain radiographs of a hand or foot

* To meet the criteria of the Classification of Psoriatic Arthritis (CASPAR) Study Group, a patient must have inflammatory articular disease (joint, spine, or enthesial) with ≥3 points from 5 categories. RF = rheumatoid factor.
† Current psoriasis is assigned a score of 2; all other features are assigned a score of 1.

polyarticular and symmetric over time. The pathophysiologic features of joint disease in PsA have recently been reviewed (5).

TENDER AND SWOLLEN JOINT ASSESSMENT

Description

Purpose. Joints are palpated for the purpose of determining if they are tender and/or swollen, the latter implying the presence of active synovitis, and both implying the presence of inflammation.

Content. Joints are assessed for tenderness and swelling. The European League Against Rheumatism (EULAR) manual of joint examination in RA (11) demonstrates appropriate examination technique. A rule of thumb is to apply ~4 kg/cm^2 of pressure (enough to blanch the tip of the examiner's fingernail) at the joint line. Joints assessed include the distal interphalangeal (DIP), proximal interphalangeal (PIP), and metacarpophalangeal joints of the hands; the wrist, elbow, shoulder, acromioclavicular, sternoclavicular, temporomandibular, hip, knee, ankle, and midtarsal joints; and the metatarsophalangeal and PIP joints of the feet.

Number of items. For clinical trials, a 68 tender and 66 swollen joint count, including the DIP joints of the hands and excluding hips for swelling, is recommended (7,12,13). The DIP joints are included because of their common involvement in PsA, unlike RA. There is also a greater tendency for more asymmetric and oligoarticular joint involvement in PsA than in RA. The DIP joints of the toes are not included because these may be difficult to evaluate reliably, and addition of these has not been demonstrated to improve performance characteristics of joint scoring systems (14). Although the 28 joint count, as used in RA, has been found to have good performance characteristics in a study of PsA phase II anti–tumor necrosis factor randomized controlled trials (RCTs) (14), it has not been recommended as an entry criteria or primary end point for RCTs because of the potential to underassess disease in the lower extremity and DIP joints, which would have resulted in 20% of patients being excluded if used to determine study eligibility (14). However, it should be noted that a retrospective analysis of the phase III infliximab trial in PsA demonstrated good performance characteristics of the simplified joint counts evaluated, including the 28 used in the Disease Activity Score in 28 joints (DAS28), a 32-joint count including the DIPs of the hands and excluding the elbows and shoulders, and a 36-joint count including the DIPs of the hands and ankles and excluding the shoulders, compared to the 68/66-joint count (15). Furthermore, in this study, only 6.5% of patients would have been excluded if these joint counts had been used to determine eligibility for trial enrollment.

Response options/scale. As in RA, the convention is to count the presence or absence of tenderness and swelling and not grade severity. Unlike RA, involvement of DIP joints is common in PsA.

In RCTs, the tender and swollen joint counts are reported separately and are used to determine the American College of Rheumatology (ACR) response and the DAS and DAS28 disease activity and EULAR response criteria, as described for RA (see article on rheumatoid arthritis disease activity), as well as the Psoriatic Arthritis Response Criteria (PsARC) (12,13), discussed below.

Recall period for items. Current presence or absence of tenderness/swelling.

Examples of use. The tender and swollen joint count is used in all clinical trials of PsA: Mease PJ. Psoriatic arthritis: update on pathophysiology, assessment and management. Ann Rheum Dis 2011;70 Suppl:i77–84 (5).

Mease PJ, Antoni CE. Psoriatic arthritis treatment: biological response modifiers. Ann Rheum Dis 2005;64 Suppl:ii78–82 (8).

Mease PJ. Psoriatic arthritis: pharmacotherapy update. Curr Rheumatol Rep 2010;12:272–80 (9).

Practical Application

Method of administration. Physical examination of 68 joints. Results are collected on a simple score sheet, on paper, or electronically.

Scoring. Presence or absence of tenderness and swelling.

Score interpretation. Used in composite measures of arthritis such as the ACR score, DAS scoring systems, or PsARC, as well as emerging composite scoring systems (see below).

Respondent burden. Minimal.

Administrative burden. Minimal; takes ~2 minutes to complete with assistant to record, or slightly longer if no assistant is available.

Psychometric Information

Method of development. Historically developed for RA assessment.

Reliability. Regarding reliability in RA, see article on rheumatoid arthritis disease activity. Regarding PsA, in a reliability exercise involving 20 experts in PsA and ankylosing spondylitis (AS), examining 10 patients with PsA and 10 with AS, the intraclass correlation coefficient (ICC) for tender joint count in the PsA patients was 0.78 (95% confidence interval [95% CI] 0.61, 0.93) and for swollen joint count was 0.50 (95% CI 0.27, 0.78) (16). The greater variability in swollen joint count may have been partly due to minimal prestudy standardization of swollen joint assessment. In a subsequent exercise involving 10 rheumatologists and 9 dermatologists examining 20 PsA patients in a Latin square design, the ICC for tender joint count among the rheumatologists was 0.81 (95% CI 0.68, 0.91) and for swollen joint count was 0.42 (95% CI 0.23, 0.65) (17). The analogous values among the dermatologists were 0.73 (95% CI 0.56, 0.86) and 0.31 (95% CI 0.23, 0.57), respectively.

Ability to detect change. The responsiveness of the 68/66 tender/swollen joint count has been demonstrated quantitatively in clinical trials of infliximab and etanercept in PsA. In a phase II trial of infliximab in PsA, the standardized response mean of 68 tender joint count in the treatment arm was −1.14 and in placebo was 0.07, for a t value of 6.0 (14). In a similar analysis of a phase II trial of etanercept, these values were −1.25, 0.53, and 6.6, respectively (14). In the same trials, the values for 66 tender joint count were −1.15 and −0.11, respectively, with a t value of 5.0 in the infliximab trial, and −1.53 and −0.20, respectively, with a t value of 5.4 (14). By comparison, the t values of the 28 tender/swollen joint counts were 5.4 and 3.8, respectively, in the infliximab trial and 5.8 and 4.1, respectively, in the etanercept trial (14). By further comparison, the t values of the 78/76 tender/swollen joint counts in the etanercept trial were 6.6 and 4.4, respectively (14).

Validity. Validation of tender and swollen joint count in PsA has not been performed.

Critical Appraisal of Overall Value to the Rheumatology Community

Joint assessment for tenderness and swelling plays a similar role as it does in RA, i.e., as a marker for presence of inflammation in joints. Because of the involvement of DIP joints and the tendency for PsA to be more asymmetric and oligoarticular, an expanded joint count of 68 tender and 66 swollen joints is recommended to more accurately assess the total burden of joint involvement. It appears that although tender joint count appears to be reliably assessed, there are still challenges in interrater reliability in swollen joint counts. In a study using dolorimetry, it has been suggested that patients with PsA display less tenderness with joint pressure than patients with RA (18).

SKIN ASSESSMENT

Psoriasis lesions may occur virtually anywhere on the skin, but are most commonly found on extensor surfaces and in the scalp. In the most common form of psoriasis, plaque psoriasis or psoriasis vulgaris, the lesions have variable degrees of erythema, induration, and scale. Most trials involving psoriasis patients are restricted to patients with this variant. Other less common variants include guttate, erythrodermic, and pustular psoriasis. Separate instruments are used to measure nail changes. The following are commonly-used psoriasis measures in psoriatic arthritis (PsA) randomized controlled trials (RCTs) and registries to assess plaque psoriasis. Although most expertly performed by dermatologists, they can be performed adequately by rheumatology clinicians trained in their use (17).

The measure most commonly used as a primary outcome measure in psoriasis trials, the Psoriasis Area and Severity Index (PASI) (19), is typically a secondary measure in PsA clinical trials. Because its performance characteristics are diminished in patients with low lesional burden, it is not typically calculated in patients with <3% body surface area (BSA) involvement with psoriasis lesions. Therefore, not all patients in an RCT will be PASI measurable. To account for this and insure some psoriasis measurement in all patients, some trials require the presence of at least 1 measurable "target lesion" of at least 2 cm in diameter. Measurable implies that the lesion is not in the scalp or groin. A BSA score (20,21) of lesional involvement and the Physician Static Global Assessment (PSGA or PGA) (22) are 2 other commonly used measures in PsA trials. Other psoriasis measures, such as the National Psoriasis Foundation Psoriasis Score (23), the Lattice System PGA (22), or the Copenhagen Psoriasis Severity Index (CoPSI) (24) scoring systems, have not been used in PsA trials or registries and therefore are not further commented upon here. These and other instruments that have been developed for psoriasis trials should be considered in the future for use in PsA trials if they show superior psychometric properties to or greater feasibility than the PASI, as has been suggested for the CoPSI (24).

The most basic assessment of psoriasis lesional burden is the BSA score (20,21). The typical method to assess BSA is to consider the surface area of the patient's handprint (palm and fingers) as representing 1% of the body's surface area. The clinician then estimates how many "handprints" would be filled by the summated lesions on the person's body. Although planimetric study suggests that the area of a flat closed hand is 0.70−0.76% of the BSA, the handprint rule has become accepted as the standard approach to estimation of BSA (25).

PSORIASIS AREA AND SEVERITY INDEX (PASI)

Description

Purpose. To provide quantitative assessment of psoriasis lesional burden based on the amount of BSA involved and degree of severity of erythema, induration, and scale, weighted by body part.

Content. The PASI was developed within a clinical trial and measures both surface area and lesional severity of psoriasis (19).

Number of items. 4 items (surface area, severity of erythema [redness], induration [thickness], and desquamation [scale]) evaluated for 4 body areas (head, trunk, and upper and lower extremities).

Response options/scale. The head, upper extremities, lower extremities, and trunk are assessed separately and then combined using weighting based on the surface area represented by each area (head = 0.1, upper extremities = 0.2, trunk = 0.3, and lower extremities = 0.4). The degree of erythema, induration, and scale in each area is judged on a 0−4 scale, the sum of which represents disease severity. The area of involvement of each area is graded from 0−6, depending on the estimated percentage of lesional area (0 = 0%, 1 = <10%, 2 = 10–29%, 3 = 30−49%, 4 = 50−69%, 5 = 70−89%, and 6 = 90–100%). These body scores are multiplied by the disease severity score and the weighting for each body area, yielding a score between 0 and 72. In trials, PASI calculators are supplied to facilitate ease of scoring.

Recall period for items. Current evaluation.

Examples of use. Widely used in clinical trials of psoriasis and PsA. Not typically used in clinical practice because of complexity.

Practical Application

How to obtain. The PASI can be obtained online at http://www.dermnetnz.org/scaly/pasi.html, as well as from the primary article on its development (19). Training videos have also been developed. The Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (online at http://www.grappanetwork.org) has developed such a video for the purpose of teaching skin and nail assessment to clinicians and trainees, and for use in the performance of clinical trials.

Method of administration. Physical examination; can be recorded on paper or entered into a computer or calculator instrument.

Scoring. The body is divided into 4 regions: head and neck (H; 10% of a person's skin), upper extremities (A; 20%), trunk (T; 30%), and lower extremities (L; 40%). Each of these areas is scored individually, and then the 4 values are combined for the total PASI score. For each body region, the percentage of area of skin involved is estimated and then assigned a grade from 0–6. Grade 0 is assigned for 0% of the area involved, grade 1 for <10%, grade 2 for 10–29%, grade 3 for 30–49%, grade 4 for 50–69%, grade 5 for 70–89%, and grade 6 for 90–100%. Within each region, the severity of psoriasis is estimated by erythema, induration, and desquamation. Severity parameters are measured on a scale of 0–4, from absent to very severe.

The sum of all 3 severity parameters is then calculated for each region of skin, multiplied by the area grade assigned for that body region and multiplied by weight of the respective section (0.1 for head and neck, 0.2 for upper extremities, 0.3 for body, and 0.4 for lower extremities).

$$\mathrm{PASI} = 0.1 \times (\mathrm{EH} + \mathrm{IH} + \mathrm{DH}) \times \mathrm{AH} + 0.2 \times (\mathrm{EA} + \mathrm{IA} + \mathrm{DA}) \times \mathrm{AA} + 0.3 \times (\mathrm{ET} + \mathrm{IT} + \mathrm{DT}) \times \mathrm{AT} + 0.4 \times (\mathrm{EL} + \mathrm{IL} + \mathrm{DL}) \times \mathrm{AL}$$

where E = erythema, I = induration, D = desquamation, and A = area.

Score interpretation. Score range is 0–72. Not reliable in patients who have <3% BSA lesional involvement and since it is rare to have a PASI score >40, nearly one-half of the scale is not used.

Respondent burden. Minimal.

Administrative burden. Takes ~5 minutes to perform with an assistant recording, or slightly longer if not.

Translations/adaptations. A simplified version has been developed, but not yet used, in PsA trials (26). A patient self-administered PASI has been shown to be reliable and correlates closely with the PASI, suggesting that it can be used when a skilled evaluator is not present (27).

Psychometric Information

Method of development. The PASI score was developed empirically for a trial of a retinoid therapy of psoriasis in 1978 (19).

Acceptability. Acceptable for clinical trials but not used in clinical practice.

Reliability. Subjective scoring of erythema, induration, and scale by a single trained observer has been demonstrated to be reliable compared to objective measures such as laser Doppler flowmeter, spectrodiometer, erythema meter, and chromameter (erythema); ultrasound (induration); and optical profilometry and scanning macrophotographic densitometry (scale) (25). In a study involving experienced and inexperienced clinicians, intrarater reliability was superior in experienced versus inexperienced clinicians ($\sigma = 1.2$ versus 3.2) (22). Interrater variation was greater, again superior in experienced compared to inexperienced clinicians ($\sigma = 8.1$ versus 9.6). In a study comparing 10 rheumatologists and 9 dermatologists, all of whom had experience with PASI scoring, examining patients with PsA, the intraclass correlation coefficient for the PASI among dermatologists was 0.74 (95% confidence interval [95% CI] 0.58, 0.87) and for rheumatologists was 0.70 (95% CI 0.53, 0.85). This suggested that the instrument could be applied reliably by trained clinicians in both specialties.

Validity. Content validity is based on the fact that the PASI measures objective skin lesion parameters of severity. However, the PASI does not include a fuller set of elements considered important to patient's assessment of severity such as "embarrassment over appearance" and itching (25). The lack of a gold standard measure of psoriasis severity limits the ability to establish criterion validity (25).

Ability to detect change. The PASI score has been used widely in clinical trials of psoriasis and PsA and demonstrates excellent ability to detect change and discriminate from placebo.

Critical Appraisal of Overall Value to the Rheumatology Community

There is poor sensitivity to change and responsiveness in mild psoriasis, so application is restricted to those with ≥3% BSA, and since it is rare to encounter a patient with a PASI score >40, nearly one-half of the range of the scale is unused. The features of erythema and scale may vary with changes in temperature and humidity and the use of emollients, instituting variables unrelated to an interventional therapy. The PASI instrument is also impractical to use in clinical practice (25). Despite these limitations, the PASI remains the most commonly used quantitative instrument to assess psoriasis in clinical trials. A standard threshold to report efficacy in a clinical trial is the PASI75 response, i.e., the percentage of patients achieving at least 75% improvement in the PASI score. This benchmark was established in a meeting between the Food and Drug Administration and the Dermatology Advisory Council in 1998. Some have criticized that this is too stringent a benchmark, in that many patients do not seek to change therapy until below a PASI50 response; the PASI50 is also associated with significant improvement in quality of life and is discriminant in clinical trials (28). However, despite these points, the PASI75 has remained the benchmark. As more effective therapies have emerged recently, PASI90 response is also measured.

The target lesion score has been used in PsA clinical trials to allow assessment of at least 1 psoriatic lesion in patients, since patients with low BSA involvement with psoriasis are not reliably measured by the PASI score. An evaluable lesion (not in the scalp, groin, or axilla) of at least 2 cm in diameter is serially evaluated for change in size (diameter) and degree of erythema, induration, and scale on a 0–3 scale of severity.

PSGA or PGA

The PGA is an overall assessment of the patient's skin lesions, on a scale of 7 descriptors, in which 0 = clear and 6 = very severe (22). It is less quantitative than the PASI but is simpler to use and is widely recognized and accepted by dermatologists.

NAIL ASSESSMENT

There is evidence of nail disease in up to 50% of patients with psoriasis and up to 80% of patients with psoriatic arthritis (PsA) (17). Characteristic nail changes involving the nail matrix include pitting, leuconychia, lunular red spots, and nail plate crumbling, whereas changes in the nail bed yield onycholysis and subungual hyperkeratosis, "oil-drop change," salmon spots, or splinter hemorrhages. The Nail Psoriasis Severity Index (NAPSI) is the most comprehensive assessment of nail disease used in psoriasis clinical trials (29). In this system, the nail is divided into 4 quadrants and 1 point is awarded if there is any finding of nail matrix and 1 point is awarded for nail bed change that is seen, per quadrant, or 0–8 per nail. This yields a potential total score of 80 if just the fingers are used and 160 if the toes are included. The original study describing the instrument showed good reproducibility (21) and a subsequent study showed good interrater reliability (30). This instrument is routinely used in psoriasis clinical trials. A modification of this system (mNAPSI) is a shorter and more feasible scoring system that has demonstrated excellent interrater reliability (31) and has been used in PsA clinical trials.

NAIL PSORIASIS SEVERITY INDEX (NAPSI)

Description

Purpose. To develop an objective reproducible tool for scoring nail psoriasis.

Content. Each nail is scored by the presence or absence of nail bed psoriasis and nail matrix psoriasis. Nail bed psoriasis includes onycholysis (separation of the nail bed), splinter hemorrhages (small, dark brown, linear marks under the nail), hyperkeratosis (thickened nail keratin), and oil-drop dyschromia (reddish-brown discoloration under the nail plate), while nail matrix psoriasis includes pitting (sharply defined depressions in the nail surface), leukonychia (white spots in the nail plate), crumbling, and red spots in the lunula.

Number of items. Evaluation of the nail bed and nail matrix are performed for each nail.

Response options/scale. Each fingernail is divided into 4 quadrants. For nail bed psoriasis, if no nail bed features are present, a score of 0 is assigned. A score of 1 is assigned if nail bed features are present in 1 quadrant of the nail, 2 if present in 2 quadrants, 3 if present in 3 quadrants, and 4 if present in 4 quadrants.

For nail matrix psoriasis, if no nail matrix features are present, a score of 0 is assigned. A score of 1 is assigned if nail matrix features are present in 1 quadrant of the nail, 2 if present in 2 quadrants, 3 if present in 3 quadrants, and 4 if present in 4 quadrants. The nail bed score and nail matrix score are added together to produce a total score for each nail, ranging from 0–8.

Recall period for items. Current.

Examples of use. Used in psoriasis randomized controlled trials (RCTs).

Practical Application

How to obtain. Available from the original article (29).

Method of administration. Physical examination, recorded on paper or electronically.

Scoring. Each nail has a possible score of 0–8, with a total possible score of 0–80 for fingernails, or 0–160 if toenails are included.

Score interpretation. Higher scores represent worse nail disease.

Respondent burden. Minimal.

Administrative burden. Takes 5–10 minutes to complete scoring, depending on amount and severity of nail disease.

Translations/adaptations. An abbreviated "target nail" version of the NAPSI is described in the original article (29). In this adaptation, 1 nail is selected, and the presence or absence of 8 parameters (onycholysis, splinter hemorrhages, hyperkeratosis, oil-drop dyschromia, pitting, leukonychia, crumbling, and red spots in the lunula) is assessed in 4 quadrants, yielding a total possible score of 0–32.

Psychometric Information

Acceptability. Acceptable for use by dermatologists in psoriasis RCTs but in PsA, either the mNAPSI or the single target nail NAPSI are utilized.

Reliability. An informal assessment of the NAPSI in its original development study, involving 37 dermatologists, demonstrated good agreement between evaluators (29).

Validity. Not validated in psoriasis or PsA. A meta-analysis of nail assessment in psoriasis RCTs has reviewed the NAPSI and other measures and identified the need for validation (32).

Ability to detect change. The NAPSI has shown responsiveness and discrimination in psoriasis RCTs. The target nail NAPSI showed responsiveness in a placebo-controlled study of golimumab in patients with PsA (33).

Critical Appraisal of Overall Value to the Rheumatology Community

The NAPSI is a detailed and highly quantitative instrument used in psoriasis RCTs but not in practice, and often a simpler measure, such as the mNAPSI or nail visual analog scale (VAS), is used in PsA clinical trials, wherein not all examiners are necessarily dermatologists, may be utilized.

MODIFIED NAIL PSORIASIS SEVERITY INDEX (MNAPSI)

Description

Purpose. To develop a nail scoring method that is simpler and more reliable than the NAPSI.

Content. For each fingernail, 7 groups of features are evaluated: pitting, onycholysis and oil-drop dyschromia, nail plate crumbling, leukonychia, splinter hemorrhages, hyperkeratosis, and red spots in the lunula. Pitting, onycholysis and oil-drop dyschromia, and crumbling (including fragmentation and horizontal ridging of the nail bed) are graded from 0–3 in severity. Leukonychia, splinter hemorrhages, hyperkeratosis, and red spots in the lunula are graded as either present or absent.

Number of items. 7 groups of features are evaluated for each fingernail.

Response options/scale. Onycholysis and oil-drop dyschromia are considered together. If no part of the nail has onycholysis or dyschromia, a score of 0 is assigned. If ≤10% of the nail has onycholysis or dyschromia a score of 1 is assigned, if 11–30% is involved a score of 2 is assigned, and if >30% is involved a score of 3 is assigned.

Pitting is scored by the number of pits present in the nail. Only pits distinctly separate from nail plate crumbling are scored. A nail with no pits is assigned a score of 0, a nail with 1–10 pits is assigned a score of 1, 11–49 pits is assigned a score of 2, and ≥50 pits is assigned a score of 3.

Crumbling may be associated with pitting. If no crumbling is present, the nail is assigned a score of 0. If crumbling is present in 1–25% of the nail, a score of 1 is assigned. If 26–50% is involved, a score of 2 is assigned. If >50% is involved, a score of 3 is assigned.

Leukonychia, splinter hemorrhages, hyperkeratosis, and red spots in the lunula are scored only by their presence or absence. A score of 1 indicates the presence of a feature, and a score of 0 indicates absence.

Recall period for items. Current.

Examples of use. Used in PsA clinical trials.

Practical Application

How to obtain. Measure is available from the original article (31).

Method of administration. Physical examination, recorded on paper or electronically.

Scoring. 0–13 per nail and 0–130 for all fingernails.

Score interpretation. Higher scores represent worse nail disease.

Respondent burden. Minimal.

Administrative burden. Scoring takes <5 minutes to perform.

Psychometric Information

Method of development. The mNAPSI was developed by rheumatologists with assistance from dermatologists as a measure simpler than the NAPSI for clinical trials.

Acceptability. Acceptable.

Reliability. Excellent intraobserver agreement has been demonstrated among PsA patients, with an intraclass correlation coefficient of 0.92 (95% confidence interval 0.87, 0.97) (31).

Validity. The mNAPSI retains the content validity of the original NAPSI in that it retains all clinically relevant aspects of psoriatic nail disease (30). A significant correlation ($P < 0.05$) has been found between mNAPSI scores and several other clinical measures of PsA (including physician global PsA disease severity VAS, swollen joint count, tender joint count, and patient global nail severity VAS), providing construct validity (31).

Ability to detect change. Responsiveness is currently being assessed in PsA clinical trials.

Critical Appraisal of Overall Value to the Rheumatology Community

This system does away with quadrant analysis and is simpler to perform, and is therefore more practical for clinical trials, and demonstrates excellent intra- and interrater reliability (31).

Nail VAS

A simple nail VAS scoring system or overall assessment of mild/moderate/severe has also been employed in PsA trials in order to gain an impression of therapeutic effect.

ENTHESITIS

Enthesitis is characterized by inflammation at sites of tendon, ligament, and joint capsule fiber insertion into bone, and is considered a pathophysiologically important aspect of psoriatic arthritis (PsA), as well as other spondylarthritides (SpA) (34). Recent registry and clinical trial patient sets have found enthesitis in approximately 30–50% of PsA patients (35). Although classically depicted involving the Achilles tendon and plantar fascia insertion sites, enthesitis can involve many parts of the body, including periknee, pelvis, spine (vertebral ligament insertion), rib cage, shoulder, and elbow. Several enthesitis scoring measures have been developed, some originally developed in patients with ankylosing spondylitis (AS). All involve a standard palpation approach, i.e., applying ~4 kg/cm^2 of pressure (enough to blanch the tip of the examiner's fingernail) and ascertaining the presence/absence and, in some indices, severity of tenderness.

MANDER/NEWCASTLE ENTHESITIS INDEX (MEI)

Purpose. The MEI was originally developed to assess all clinically accessible entheses potentially involved in AS.

Method of development. Based on clinical experience, the investigators identified a large number of potentially involved enthesial sites. After removing sites that did not produce tenderness on palpation in any of 19 study patients, the instrument specified 66 sites for assessment (36).

Scoring. A scoring system based on the patient's response to palpation over the entheses is rated from 0–3 (where 0 = no pain, 1 = mild tenderness, 2 = moderate tenderness, and 3 = wince or withdraw). A maximum total score of 90 is possible (36).

Examples of use. The MEI has not been used in randomized controlled trials (RCTs) because of burden of administration and concern about reliability.

Critical Appraisal of Overall Value to the Rheumatology Community

The instrument has been criticized for the large number of sites examined, rendering it too time consuming for use in clinical trials, as well as overlap of many sites with fibromyalgia tender point sites. Further, the 0–3 scoring system could contribute to greater inter- and intrarater inconsistency. It has never been used in an RCT and therefore has not been evaluated for reliability or responsiveness. However, it is often referred to for the purpose of describing the overall set of potential enthesis sites from which other measures have derived their simpler version. Indeed, the Maastricht scoring system (see below) was derived from the MEI and in the process, a validation exercise for the MEI was performed (16).

LEEDS ENTHESITIS INDEX (LEI)

Description

Purpose. To assess enthesitis in patients with PsA. Whereas other enthesitis measures described here were developed and/or validated in patients with AS, the LEI was developed specifically for PsA (37).

Content. Enthesial sites include the bilateral lateral epicondyles, medial femoral condyles, and Achilles tendon insertions.

Number of items. 6 enthesial sites.

Response options/scale. Presence or absence of tenderness.

Recall period for items. Current.

Examples of use. Used in several PsA trials being conducted currently.

Practical Application

How to obtain. Description of sites can be found in the original article (37) and examination technique is present on the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) web site (http://www.grappanetwork.org).

Method of administration. Physical examination; results can be recorded on paper document or electronically.

Scoring. Tenderness on examination is recorded as either present (1) or absent (0) for each of the 6 sites, for an overall score range of 0–6.

Score interpretation. Higher count represents greater enthesitis burden.

Respondent burden. Minimal.

Administrative burden. Takes ~30 seconds to complete.

Psychometric Information

Method of development. Similar to the methodology used to develop the Maastricht Ankylosing Spondylitis Enthesis Score (MASES; see below), the 6 sites of the measure were selected based on a stepwise data reduction to identify those sites most commonly involved.

Acceptability. Acceptable.

Reliability. In the study involving comparison of measures in the assessment of AS and PsA spondylitis patients, the LEI demonstrated an intraclass correlation coefficient (ICC) of 0.81 (95% confidence interval [95% CI] 0.65, 0.94) (16).

Validity. In a study comparing the LEI, MEI, MASES, modified Spondyloarthritis Research Consortium of Canada (SPARCC; 8 sites), and Major indices in PsA patients commencing disease-modifying therapy, clinical parameters of disease activity correlated most consistently with the LEI (37).

Ability to detect change. In the study described above, the greatest effect size at 6 months was demonstrated with the LEI and modified SPARCC, moderate change with the Major, and small for the MASES and MEI (37). In this same exercise, the LEI showed the least floor effect (scoring 0 when MEI was >0) of any of these indices (37).

Critical Appraisal of Overall Value to the Rheumatology Community

In an open-label longitudinal treatment study including several enthesial measures, the LEI showed the closest correlation with other disease activity measures, was responsive, and showed the least floor effect, i.e., indicating its ability to identify the majority of PsA patients with enthesitis (37).

SPONDYLOARTHRITIS RESEARCH CONSORTIUM OF CANADA (SPARCC)

Description

Purpose. To assess enthesitis in patients with SpA. The SPARCC created a measure for enthesitis in SpA in general (i.e., not limited to PsA or AS).

Content. Enthesial sites examined include the bilateral Achilles tendons, plantar fascia insertion at the calcaneus, patellar tendon insertion at the base of the patella, quadriceps insertion into the superior border of the patella, supraspinatus insertion into the greater tuberosity of the humerus, and medial and lateral epicondyles.

Number of items. 16 enthesial sites.

Response options/scale. Presence or absence of tenderness.

Recall period for items. Current.

Practical Application

How to obtain. List of sites is in the original article (38). Examination method may be viewed online at www.arthritisdoctor.ca and is also present on the GRAPPA web site (http://www.grappanetwork.org).

Method of administration. Physical examination; results can be recorded on paper document or electronically.

Scoring. Tenderness on examination is recorded as either present (1) or absent (0) for each of the 16 sites, for an overall score range of 0–16.

Score interpretation. Higher count represents greater enthesitis burden.

Respondent burden. Minimal.

Administrative burden. Takes ~2–5 minutes to complete.

Translations/adaptations. Modified versions with fewer but more commonly involved sites (6 and 8 sites) showed greater responsiveness and were more discriminant between treatment and placebo (6 sites) (38).

Psychometric Information

Method of development. Selection of enthesitis sites was based on information from published power Doppler ultrasound in SpA patients compared to rheumatoid arthritis patients and healthy controls, and magnetic resonance imaging studies of the shoulder in AS patients. The most frequent enthesitic sites were selected (38).

Acceptability. Acceptable.

Reliability. In a study comparing enthesitis indices in patients with AS versus PsA with spondylitis, the 8-site SPARCC index showed an ICC of 0.81 (95% CI 0.64, 0.93) (16).

Validity. The instrument has not been validated in PsA. In patients with AS, substantial correlations have been observed between the SPARCC enthesitis score and the Bath Ankylosing Spondylitis Disease Activity Index, Bath Ankylosing Spondylitis Function Index, and patient global (38).

Ability to detect change. In a study of 9 patients randomized to adalimumab, a nonsignificant reduction in the SPARCC index scores was recorded at 12 weeks. The SPARCC scores had decreased further at the 24-week assessment ($P = 0.04$).

MAASTRICHT ANKYLOSING SPONDYLITIS ENTHESIS SCORE (MASES)

Description

Purpose. The original purpose was for the assessment of enthesitis in AS, and now is additionally used in PsA and SpA in general.

Content. Clinical scoring system for enthesitis in SpA, including AS and PsA. Enthesial sites assessed include the bilateral first costochondral joints, seventh costochondral joints, posterior superior iliac spines, anterior superior iliac spines, iliac crests, proximal insertion of Achilles tendons, and the fifth lumbar spinous process.

Number of items. 13 enthesial sites.

Response options/scale. Presence or absence of tenderness.

Recall period for items. Current.

Endorsements. Recommended by the Assessment of SpondyloArthritis international Society for use in randomized controlled trials of AS and SpA.

Examples of use. Used in AS clinical trials, emerging trials with the recently adopted axial SpA criteria, and PsA trials.

Practical Application

How to obtain. List of entheses in the original article (39) and examination technique is demonstrated on the GRAPPA web site (http://www.grappanetwork.org).

Method of administration. Physical examination; results can be recorded on paper document or electronically.

Scoring. Tenderness on examination is recorded as either present (1) or absent (0) for each of the 13 sites, for an overall score range of 0–13 (39).

Score interpretation. Higher scores reflect greater enthesitis burden.

Respondent burden. Minimal.

Administrative burden. Completion by health professional takes 2–5 minutes.

Psychometric Information

Method of development. Recognizing that the MEI was too lengthy for use in clinical trials, Heuft-Dorenbosch and colleagues employed the MEI in AS patients over 2 years, and selected the 13 most specific and sensitive sites from that index to constitute the MASES (39).

Acceptability. Acceptable. The MASES correlates well with the MEI, and the reduction in site number and removal of intensity grading yields a more practical instrument.

Reliability. In a study in which several enthesitis indices were compared in the evaluation of patients with AS or PsA with spondylitis, moderate intraobserver agreement was demonstrated among PsA patients, with an ICC of 0.56 (95% CI 0.34, 0.82). The ICC for the MASES was greater in patients with AS than in PsA (16).

Validity. The instrument has not been validated in PsA.

Ability to detect change. Discrimination and responsiveness have been demonstrated in a trial of golimumab in PsA (33) as well as multiple studies in AS.

Berlin (Major)

This is a 12-site enthesitis index (40) (Table 2) used in studies of AS, and was also evaluated in the International Spondyloarthritis Interobserver Reliability Exercise (INSPIRE) trial, although it showed lower ICC values than the Leeds and SPARCC instruments in PsA patients (16). This instrument has not been used in PsA trials.

San Francisco

This is a 14-site enthesitis index (41) (Table 2) employed in trials of AS that, like the Major, showed lower ICCs in PsA patients in the INSPIRE study and has not been used in PsA trials (16).

4 Point

The 4 point enthesitis measure includes both Achilles tendon and plantar fascia insertions and may be graded as present or absent or scored on a 0–3 scale of severity. This

Table 2. Enthesial sites assessed in outcome measures for enthesitis (16)*

		Major		San	PEST	4
	MASES	(Berlin)	SPARCC	Francisco	(Leeds)	point
C1/C2				X		
C7/T1				X		
T12/L1				X		
First costochondral	R, L					
Seventh costochondral	R, L					
Supraspinatus insertion			R, L			
Lateral epicondyle humerus			R, L		R, L	
Medial epicondyle humerus			R, L			
Posterior superior iliac spine	R, L					
Anterior superior iliac spine	R, L			R, L		
Iliac crest	R, L	R, L				
Fifth lumbar spinous process	X			X		
Ischial tuberosity				R, L		
Proximal Achilles	R, L	R, L	R, L	R, L	R, L	R, L
Greater trochanter		R, L	R, L	R, L		
Medial condyle femur		R, L			R, L	
Lateral condyle femur		R, L				
Insertion plantar fascia		R, L	R, L	R, L		R, L
Quadriceps insertion patella			R, L			
Inferior pole patella			R, L			
Tibial tubercle			R, L			

* MASES = Maastricht Ankylosing Spondylitis Enthesis Score; SPARCC = Spondyloarthritis Research Consortium of Canada; PEST = Psoriasis Epidemiology Screening Tool; X = single site present, not bilateral; R = right; L = left.

measure has shown discrimination and responsiveness in 2 infliximab trials in PsA and the etanercept trial known as Psoriasis Randomized Etanercept Study in Subjects with Psoriatic Arthritis (5,8,9); however, it did not perform as well as measures with a greater number of sites in a comparative study (16).

DACTYLITIS

Dactylitis, or "sausage digit," is characterized by swelling of an entire finger due to synovitis, tenosynovitis, enthesitis, and soft tissue edema (12,13,42). Its presence helps distinguish psoriatic arthritis (PsA) from other forms of arthritis, and is found in 16–48% of reported PsA cases (43). In PsA clinical trials over the past decade (5,8,9,44), dactylitis has been assessed by having the investigator examine each finger and determine if it is swollen or not (12,13,45). On occasion, a severity score of 0–3 (where 0 = no swelling or pain and 3 = severe swelling and pain) has been utilized as well. Both the methods of simple count and scoring have demonstrated responsiveness and discrimination in anti–tumor necrosis factor trials (5,8,9). Recently, a more quantitative dactylitis measure, the Leeds Dactylitis Index (LDI), has been developed and is described below.

LEEDS DACTYLITIS INDEX (LDI)

Description

Purpose. To quantitatively measure dactylitis. The LDI, using a measurement instrument known as a dactylometer, has been recently developed but not yet used in a randomized controlled trial (RCT) (43). In this method, circumference of the affected fingers, circumference of contralateral fingers, and tenderness of affected fingers are all assessed for a total score.

Content. Evaluation of finger size and pain to assess for the presence of dactylitis.

Number of items. Evaluation of each of 20 fingers for size and tenderness.

Response options/scale. See below.

Recall period for items. Current.

Examples of use. Currently being used in PsA RCTs.

Practical Application

How to obtain. The measure is available in the original article (43). The tool developed for measuring digital circumference is available online at www.mie-uk.com.

Method of administration. The clinician marks which fingers are affected on a diagram displaying fingers and toes. Circumferences of the affected and contralateral fingers are then measured around the proximal phalanx, as close as possible to the web space, using either a measuring tape or a precalibrated loop. The clinician then squeezes the affected fingers with moderate pressure and documents the patient's response: 0 = no tenderness, 1 = tender, 2 = tender and winces, and 3 = tender and withdraws (43).

Scoring. The ratio of circumference between an affected finger and the contralateral unaffected finger is recorded. If both sides are affected, the circumference of the affected finger is compared to normative data supplied in a table. The tenderness score (0–3) for a finger with dactylitis is recorded, and a total score is generated for each finger. If multiple fingers are affected, each score is added together to produce a total for the patient (38). A difference in digital circumference of ≥10% is used to define a finger with dactylitis.

Score interpretation. A higher score is associated with worse dactylitis (43).

Respondent burden. Minimal burden to the patient; no discomfort.

Administrative burden. Assessment and scoring time by the clinician depends on the number of affected fingers; however, total administration should take <10 minutes (43).

Translations/adaptations. A later modification of the LDI (referred to as the LDI basic) replaced the original tenderness grading (0–3) with a binary score reflecting either the presence or absence of tenderness (1 or 0, respectively) (45).

Psychometric Information

Method of development. Developed in a patient study specific for the development of this measure (43).

Acceptability. Acceptable.

Reliability. In a study involving 20 rheumatologists expert in PsA and ankylosing spondylitis who examined 20 patients, overall interobserver scores indicated a strong agreement, with an intraclass correlation coefficient of 0.70 (95% confidence interval 0.49, 0.89) (16).

Validity. Discrimination capability has not yet been assessed and the measure has not been validated.

Ability to detect change. Responsiveness to change has been demonstrated in a small open study of treated PsA patients (45).

Critical Appraisal of Overall Value to the Rheumatology Community

The strength of this method is its quantitative objectivity. A weakness is that it takes more time to perform than an observational count. When matching fingers are involved, reference values gathered by Helliwell and colleagues in 2005 are used to calculate the dactylitis score. While these values are grouped by sex, they do not control for other variables such as age or body mass index (43).

SPINE ASSESSMENT

The prevalence and impact of spine disease in psoriatic arthritis (PsA) has not been as well characterized as more peripheral manifestations of the disease. Classified as a spondylarthritis based on overlapping clinical and pathologic features with ankylosing spondylitis (AS), spinal inflammation does occur in up to one-half of PsA patients. Manifestations include sacroiliitis, spinal joint and enthesial inflammation, and ankylosis in the form of bridging syndesmophytes. The spinal manifestations of PsA tend to be less severe than those seen in AS. Because spine involvement tends to be mild and inconstant, it has not been systematically assessed in clinical trials of PsA. Rather, it appears that investigators, clinicians, and regulators tend to adopt the research findings of AS and extrapolate to PsA in the spine in the absence of specific clinical trial data from PsA patients.

Commonly used measures in AS trials include the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) (46), Bath Ankylosing Spondylitis Function Index (BASFI) (47), and Bath Ankylosing Spondylitis Metrology Index (48). The BASDAI is a set of 6 visual analog scale (VAS) patient questionnaires regarding fatigue, pain, and stiffness. A study of the BASDAI in PsA showed a high correlation with a VAS of overall arthritis activity, but the correlation was similar in patients with a greater amount of axial disease as compared to patients with a greater amount of peripheral disease (49). Also, it did not correlate with the physician's perception of disease activity or with treatment decisions. Therefore, the BASDAI did not discriminate between axial and peripheral disease activity. Similarly, the BASFI demonstrated a high correlation with other measures of function, such as the Health Assessment Questionnaire (HAQ) disability index (DI) and Short Form 36 (SF-36), but did not discriminate between patients with predominantly axial versus peripheral disease, and therefore offers no advantage over the commonly used HAQ DI and SF-36 (50).

In a study of 10 patients with AS and 10 patients with PsA spondylitis, 10 experts in AS and 10 experts in PsA performed AS spine examinations and found there to be good performance characteristics and interobserver reliability (intraclass correlation coefficient 0.89) using standard AS metrology in both AS and PsA spondylitis patients (51). In a similar exercise, spine measurement techniques, in particular the modified Schober test, lumbar side flexion, and cervical rotation, compared favorably between patients with PsA spondylitis and AS (52).

The recently developed Ankylosing Spondylitis Disease Activity Score (53) was compared to the BASDAI in patients with PsA spondylitis. Both showed good correlation with disease activity measured by patient and physician global and both performed comparably (54). It appears that if measures of axial disease developed for AS are used in trials of PsA patients with axial involvement, the measures will be reasonably reliable, responsive, and discriminative.

PATIENT GLOBAL ASSESSMENT

The intent of the patient global assessment is to encompass not only specific disease severity and multidimensional impact, but also take into account the impact of such issues as treatment side effects. Typically, when a patient is asked to judge their global status on a visual analog scale (VAS) or Likert scale, it is in relation to the following question: "In all of the ways that your [insert disease name] affects you [today, or over the past week], how are you?" American College of Rheumatology (ACR) criteria to assess rheumatoid arthritis (RA) (55) express that the patient global assessment of disease activity, by VAS score, should be answered in response to the following statement: "Considering all the ways your arthritis affects you, please mark a vertical line on the scale below to show how you are feeling today." Zero is considered to be "very good, no symptoms" and 100 is "very poor, severe symptoms." If responses are offered on a Likert scale, the same question regarding the effect of arthritis today is asked and the patient circles a number ranging from 1, denoting "very good, no symptoms and no limitation of normal activities," to 5, denoting "very poor, very severe symptoms which are intolerable and inability to carry out normal activities."

In psoriatic arthritis (PsA), the patient (and the clinician asking the question) may not understand if they should focus their thinking on joint disease, all aspects of musculoskeletal disease, skin and nail disease, or a composite of these.

Illustrating the conundrum about how patient global should be assessed in a PsA trial, 1 recent PsA trial protocol included the above sentence to describe the patient global, but with the instruction that the subject should be asked to consider both joint and skin components in their response to this statement (Mease PJ: unpublished observations), even though asking the question in this way has not been formally validated in the ACR or Disease Activity Score (DAS) criteria. In most PsA protocols, there has been no such clarification and it has been left up to the interpretation of the patient and/or investigator/coordinator.

To address this issue, the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) organized a multicenter study in which patients were asked, in variable order, to rate on a VAS scale their assessment of psoriasis and arthritis (PGA), arthritis alone (PJA), and psoriasis alone (PSA) over the past week (56). For example, the PGA question was, "In all the ways that your psoriasis and arthritis, as a whole, affect you, how would you rate the way you felt over the past week?"

PATIENT GLOBAL FOR PSORIATIC ARTHRITIS

Description

Purpose. To adequately capture the patient global experience of PsA, taking into consideration the separate domains of musculoskeletal and skin disease.

Content. Questions asked of the patient about patient global experience.

Number of items. 3 self-reported questions.

Response options/scale. The patient rates global state related to arthritis, skin, and a combination of these on VAS or numerical rating scales.

Recall period for items. Past week.

Endorsements. Endorsed by GRAPPA.

Practical Application

How to obtain. From the original article (56).

Method of administration. Self-administered.

Scoring. In addition to being scored on a VAS, the patient global assessment of arthritis has been scored on a Likert scale.

Score interpretation. The patient global assessment is used in the ACR, DAS, and Psoriatic Arthritis Response Criteria scoring systems (see below), as well as emerging composite scoring systems.

Respondent burden. Minimal.

Administrative burden. Minimal.

Psychometric Information

Method of development. Evaluation of 319 patients seen in globally distributed clinics of GRAPPA members.

Acceptability. Acceptable.

Reliability. All 3 measures, i.e., the PGA, PJA, and PSA, demonstrated good test–retest reliability with intraclass correlation coefficients 0.87 (95% confidence interval [95% CI] 0.83, 0.90), 0.86 (95% CI 0.81, 0.89), and 0.78 (95% CI 0.72, 0.83), respectively. Analysis of the GRAPPA study responses found that the PJA had a B coefficient of 0.63, while the PSA had a B coefficient of 0.30. The regression coefficient B quantifies the effect of each exposure variable by expressing the increase of the outcome variables produced by a unitary increment of the exposure variable (56). This disparity indicates that in general, the study patients considered the arthritis component of their disease to be a greater problem than the psoriasis component.

Validity. Analysis of the recent GRAPPA study responses shows a statistically significant correlation between higher values on the patient global VAS and the number of joints involved (56).

Ability to detect change. To be determined based on results of randomized controlled trials (RCTs) currently being conducted. Historically, the PJA, as used in RA, has been responsive and able to discriminate treatment from placebo.

Critical Appraisal of Overall Value to the Rheumatology Community

The recent GRAPPA study found that although the single question addressing both joint and skin disease is a reliable and responsive measure, that since some patients had divergent severity of joint and skin disease, the variable impact of the 2 major aspects of the disease could be better understood by asking all 3 questions, if feasible, in an RCT.

Since the ACR and DAS scoring systems allow for only 1 entry of patient global in the score, based on the study recently published by the GRAPPA group described above (56), it would appear reasonable to use either the dual question (PGA) or the question related to arthritis alone (PJA).

PATIENT PAIN ASSESSMENT

Assessment of pain related to arthritis is part of the core set of measures for the American College of Rheumatology score (55). Patient pain is assessed by means of a 0–100 visual analog scale (VAS), where 0 = "no pain" and 100 = "most severe pain" that day. The minimum clinically important difference of pain VAS is considered to be 10 mm (57).

PHYSICIAN'S GLOBAL ASSESSMENT OF DISEASE ACTIVITY

The physician's global assessment of disease activity is scored in an identical manner to the patient global assessment of disease activity described above, both in visual analog scale format for the American College of Rheumatology scoring system and the Likert scale format for the Psoriatic Arthritis Response Criteria. The same problem regarding whether one focuses on joint or skin activity or both applies, and in recent psoriatic arthritis study protocols, the clinician has been instructed to take into account both the severity of joint and skin disease when making the assessment (Mease PJ: unpublished observations).

PHYSICAL FUNCTION

Physical function has been reliably assessed in psoriatic arthritis (PsA) trials by the Health Assessment Questionnaire (HAQ; see below) (58). This measure contains 20 items divided into 8 domains: dressing and grooming, arising, eating, walking, hygiene, reach, grip, and common daily activities. Subjects rate the degree of difficulty they have had in the past week on a 4-point scale, ranging from 0 (no difficulty) to 3 (unable to do). The highest scores in each category are summed (0–24) and divided by the number of categories scored to yield a score from 0–3. The HAQ instrument has been most extensively utilized and validated in rheumatoid arthritis (RA). In an effort to modify this instrument to be more specific for spondylarthritides in general and PsA specifically, the HAQ for spondylarthritis (59) and the expanded HAQ (60) were developed with questions more specific for spondylarthritis and psoriasis patients. These modifications have not been found to be psychometrically superior to the original HAQ, and so have not been further used in PsA clinical trials. In a large cohort of patients in the Toronto PsA registry, the number of inflamed joints (reflecting disease activity) and deformed joints (reflecting damage) were significantly related to the HAQ score (61). The effect of disease activity on the HAQ declined with duration of disease activity, as has been noted in RA, but there was less evidence that the effect of joint damage on HAQ increased over time. A Rasch analysis showed that the Short Form 36 (SF-36) physical functioning (PF) subscale and the HAQ measure the same physical disability construct, but the SF-36 PF subscale has better distributional properties, scale length, fewer misfitting items, and less differential item functioning (62).

The minimum important difference (MID) of the HAQ, determined using an anchor-based method, calculated using the patient-rated importance of change in a trial of etanercept in PsA (63), was 0.35 (64). Calculations based on distribution-based methods are thought to be a more appropriate approach to analysis of MID than that based on patient-rated satisfaction with change, noted to be 0.3 in a preliminary abstract from this same study (65). In contrast, a study from a single center, using methods based on an overall health status anchor, found the MID to be 0.131 (66). The difference may have been partly accounted for by the difference in disability of patients in the 2 studies, with an average HAQ score of 1.16 in the former study and 0.732 in the latter, as well as trial methodology. Both of the figures derived from these 2 studies contrast with the minimal clinically important difference, a synonymous term for MID, established for RA of 0.22 (67).

In addition to the HAQ and the SF-36 PF subscale, the Disabilities of Arm, Shoulder, and Hand questionnaire has been validated for assessment of upper extremity function and inflammatory disease activity in PsA. The Arthritis Impact Measurement Scales (AIMS) and a revised form of this instrument (AIMS2) have been validated as measures of function in PsA in a PsA registry, but these measures have not been used in PsA clinical trials (68,69).

HEALTH-RELATED QUALITY OF LIFE

The most commonly used measure of quality of life (QOL) employed in psoriatic arthritis (PsA) trials is the generic QOL instrument, the Medical Outcomes Study Short Form 36 (SF-36) (70). Other commonly used measures have included the Dermatology Life Quality Index (DLQI) (71) and the EuroQol 5-domain (EQ-5D) (72). A PsA-specific instrument, the Psoriatic Arthritis Quality of Life (PsAQOL) measure, has been developed and validated.

The SF-36 is a patient questionnaire assessing 8 domains of health status: physical functioning (PF), pain, vitality, social functioning, psychological functioning, general health perceptions, and role limitations due to physical and emotional problems (70). It also can be subdivided into 2 summary scores, the physical and mental component scores. This instrument has been validated in PsA (73). All domains of the SF-36 demonstrated internal consistency reliability, with Cronbach's α exceeding 0.8. Discriminant validity was demonstrated in that the adjusted SF-36 scale scores for patients with PsA were lower than age- and sex-matched controls. Discriminant and convergent validity were demonstrated by correlation between SF-36 domains and clinical measures of function and pain and measures of disease activity and severity. A more recent study used Rasch analysis to compare the SF-36 and Health Assessment Questionnaire (HAQ) in PsA and rheumatoid arthritis (RA) (62). The Rasch model fit the SF-36 PF subscale scores better than the HAQ with good item separation in both PsA and RA and minimal floor effects in either group. Rasch analysis of the HAQ demonstrated a better span of the HAQ in RA than in PsA, and there were significant floor effects in PsA, with 30% indicating no disability and only 7% of the RA group indicating no disability. The HAQ and SF-36 PF subscale measure the same physical disability construct, and the SF-36 PF subscale has better distributional properties and scale length, fewer misfit items, and less differential item functioning than the HAQ. Another recent study assessed the scaling assumptions, internal reliability, and construct validity of the SF-36 in PsA in an Asian population in Hong Kong, further validating it (62). As previously described, the SF-36 physical component summary score demonstrated superior psychometric properties compared to the HAQ as well as to the Bath Ankylosing Spondylitis Functional Index and Dougados Functional Index (50).

DERMATOLOGY LIFE QUALITY INDEX (DLQI)

Description

Purpose. The DLQI was developed to measure the disability experienced by patients with different dermatologic conditions. The DLQI is a 10-item questionnaire used in 33 different dermatologic conditions (71). It has been validated in assessment of psoriasis and shows discrimination and responsiveness in PsA trials (5,8,9). It is often used in conjunction with the SF-36 to provide more specific assessment of the impact of skin disease.

Content. Questions assess the effect of a dermatologic condition on a patient's QOL, including impact on work, leisure activities, personal relationships, feelings of embarrassment, etc.

Number of items. 10 items.

Response options/scale. Answers are based on a 4-point Likert scale. Responses of "not at all," "a little," "a lot," and "very much" are available for each question, and correspond to scores of 0, 1, 2, and 3, respectively. A response of "not relevant" is also offered for select questions (71).

Recall period for items. 7 days.

Examples of use. Widely used in psoriasis and PsA clinical trials.

Practical Application

How to obtain. Measure is available from the original article (71).

Method of administration. Can be self-administered on paper or electronically.

Scoring. 0–3 for each question, yielding a total possible score of 0–30.

Score interpretation. Higher scores represent a greater effect on QOL.

Respondent burden. Minimal.

Administrative burden. Minimal.

Psychometric Information

Method of development. The DLQI was developed through a study of 120 dermatology patients presenting with a variety of skin conditions. Patients were asked questions about their skin disease and its impact on their lives, and a 10-item questionnaire was formulated based in their answers.

Acceptability. Acceptable.

Reliability. Reliability has been demonstrated using a 1-week test–retest method in 53 patients (71). Correlations between individual question scores were very high (γ_s = 0.95–0.98, $P < 0.001$), and the correlation between overall DLQI scores was also high ($\gamma_s = 0.99$, $P < 0.0001$).

Validity. Construct validity has been demonstrated in a comparison of scores of 200 dermatology outpatients compared to 100 control subjects. In a separate study, the DLQI was cross-validated against other QOL measures such as the SF-36 (74).

Ability to detect change. The DLQI has demonstrated sensitivity to change in 181 patients following inpatient treatment for their dermatologic conditions (75). It is used widely in psoriasis and PsA clinical trials and shows significant responsiveness.

Critical Appraisal of Overall Value to the Rheumatology Community

The DLQI is often included in PsA randomized controlled trials (RCTs), along with the HAQ, SF-36, and currently, the PsAQOL measure, partly as a measure that dermatologists can relate to when evaluating QOL effects of therapies. It is one of the measures that may be used to assess disease severity when utilizing the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis treatment recommendations grid (76).

EQ-5D

The EQ-5D is comprised of a 5-item set of health status measures and a visual analog scale (VAS), with each of the 5 health states (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) evaluated from "no problem" to "extreme problem," scored from 1–3 (72). The VAS is rated from 0 = worst imaginable health status to 100 = best imaginable health status. The EQ-5D has shown discrimination and responsiveness in PsA trials (5,8,9).

PSORIATIC ARTHRITIS QUALITY OF LIFE (PSAQOL)

Description

Purpose. PsA-specific health-related QOL instrument. The PsAQOL, derived from PsA patient interviews and evaluations, has shown reliability and construct validity (77). It is now being employed in RCTs (Mease PJ: unpublished observations) to further assess discrimination and responsiveness.

Number of items. 20 items.

Response options/scale. Questions are answered as "true" or "not true."

Recall period for items. Current.

Examples of use. Currently being employed in PsA RCTs.

Practical Application

How to obtain. Contact Stephen McKenna (e-mail: smckenna@galen-research.com).

Method of administration. Self-administered on paper or electronically.

Scoring. Each "true" response is 1 point on a 20-point scale, for a possible total score of 0–20.

Score interpretation. Higher scores indicate worse health-related QOL.

Respondent burden. 1–2 minutes to complete.

Administrative burden. Minimal.

Psychometric Information

Method of development. Qualitative interviews were conducted with 48 PsA patients, from which a 51-item questionnaire was generated. Followup surveys of 94 patients reduced the number of relevant questions to 35, and a Rasch analysis of an additional 286 patient surveys identified 20 meaningful items to include in the final PsAQOL questionnaire (77).

Acceptability. Acceptable.

Reliability. The PsAQOL has demonstrated excellent test–retest reliability (Spearman's rank correlation coefficient = 0.89) (77).

Validity. Strong correlations have been found between the PsAQOL and other comparable measures (77). Construct validity of the PsAQOL has also been confirmed in a longitudinal study of 28 patients over 6 months (78).

Ability to detect change. Responsiveness has been demonstrated at both 3-month and 6-month time points after treatment initiation in PsA patients (78).

Critical Appraisal of Overall Value to the Rheumatology Community

The PsAQOL has been rigorously developed and is specific for patients with PsA. It is now being used for the first time in RCTs, from which more will be learned about its performance characteristics.

In a series of PsA patient focus groups, transcribed texts of the discussions were divided into meaning units from which concepts were extracted and mapped using the International Classification of Functioning, Disability and Health as a frame of reference (79). Multiple commonly used measures of function and QOL, including those described herein, demonstrated significant gaps in coverage of key concepts important to patients in any one instrument, suggesting the need to use several instruments to more adequately address these concepts, as well as the need for more research on development of more comprehensive instruments to measure function and QOL.

FATIGUE

Fatigue has increasingly been recognized as an important clinical dimension by patients with rheumatic diseases (80). Fatigue has been determined to be an important clinical domain in psoriatic arthritis (PsA), independent of and not fully explained by other domains such as pain, tender and swollen joint count, patient global, and function (81). Several scales have been used to assess fatigue in rheumatic diseases, including the Multidimensional Assessment of Fatigue scale (82), the Multidimensional Fatigue Index (83), the Fatigue Severity Scale (FSS) (84), the Functional Assessment of Chronic Illness Therapy–Fatigue (FACIT-F) scale (85), the Brief Fatigue Inventory (86), the vitality scale of the Short Form 36 (70), and the visual analog scale for fatigue (87). A modified version of the FSS has been validated in PsA (88). Changes in fatigue were correlated with changes in disease activity using this instrument (89). The FACIT-F was responsive to change in a trial of adalimumab in PsA (90). The FACIT-F was validated in a Toronto PsA cohort and correlated well with the modified FSS, showing high internal consistency, test–retest reliability, and criterion and construct validity (91).

FUNCTIONAL ASSESSMENT OF CHRONIC ILLNESS THERAPY–FATIGUE (FACIT-F)

Description

Purpose. The FACIT measurement system was originally developed to assess health-related quality of life in patients with chronic illnesses (92,93). The additional questions of the FACIT-F survey were compiled to assess anemia-related fatigue (94).

Number of items. 13 items.

Response options/scale. Answers are based on a 5-point Likert scale. Responses of "not at all," "a little," "somewhat," "quite a bit," and "very much" are available for each question, and correspond to scores of 0, 1, 2, 3, and 4, respectively.

Recall period for items. 7 days.

Practical Application

How to obtain. The questionnaire can be found online at http://www.facit.org.

Method of administration. Self-administered on paper or electronically.

Scoring. Each question scores between 0 and 4, with a total score range from 0–52.

Respondent burden. Minimal.

Administrative burden. Minimal.

Psychometric Information

Method of development. Each FACIT measure is developed through interview-based item generation and survey-based item reduction (93).

Acceptability. Acceptable.

Reliability. High test–retest reliability of the FACIT-F was demonstrated in a Toronto study of 135 patients with PsA. FACIT-F surveys given to patients 1 week apart yielded an intraclass correlation coefficient of 0.95 (91).

Validity. The measure was validated in the same Toronto PsA cohort. It correlated well with the modified FSS (−0.79; 95% confidence interval −0.85, −0.72) and showed high internal consistency (Cronbach's $\alpha = 0.96$), as well as criterion and construct validity (91).

Ability to detect change. The FACIT-F was responsive to change in a trial of adalimumab in PsA (90).

Critical Appraisal of Overall Value to the Rheumatology Community

The domain of fatigue, as mentioned, is increasingly recognized as an important domain to assess in patients with inflammatory arthritis conditions. No single measure has emerged as a favored instrument, although several, as described, have been shown to be reliable, responsive, and discriminative. The area of fatigue assessment is still in evolution in PsA.

PATIENT-REPORTED OUTCOMES MEASUREMENT INFORMATION SYSTEM (PROMIS)

In 2004, the National Institutes of Health Roadmap initiative PROMIS was launched to develop a new generation of patient-reported outcome measures for chronic illness facilitated by the methods of item-response theory and computer adaptive testing. PROMIS is intended to be a publicly available updateable repository of well-calibrated items facilitating the assessment of numerous clinically relevant domains across disease states with precision and minimal patient burden (95). The intent is for patients to be able to self-report about disease activity, function, and quality of life using online questionnaires, centrally and securely stored electronically. This is a public–private partnership intended to facilitate disease state registries and clinical trials. It is anticipated that this methodology will be used in the assessment of psoriatic arthritis as well as other rheumatic diseases, hence the description in this article. The PROMIS web site is www.nihpromis.org.

IMAGING

Although it is beyond the scope of this article to describe in detail imaging assessments in psoriatic arthritis (PsA), providing a few words about the approach to imaging in PsA trials and steering the reader to reviews on this subject are appropriate. Radiographic imaging is used for diagnostic purposes and to assess joint damage. Ultrasound (US) and magnetic resonance imaging (MRI) can be used for these same purposes, and to assess for the presence of inflammation through detection of soft tissue changes such as synovitis, enthesitis, excess fluid, and increased blood flow (US power Doppler). Additionally, MRI can detect inflammatory changes in bone by detecting bone edema.

In terms of radiographic assessment, the most commonly employed methodology used in PsA clinical trials is the Sharp/van der Heijde method (hands, wrists, and feet) modified for PsA to include the distal interphalangeal joints. A detailed review of this and other radiographic methods used in PsA is provided in a review article by Mease and van der Heijde (96) and other reviews (97). This method has shown excellent response and discrimination characteristics in PsA clinical trials (5,8,9,44).

Although not conducted in PsA clinical trials as of yet, spine radiographic methods used in ankylosing spondylitis, the Bath Ankylosing Spondylitis Radiology Index and the modified Stoke Ankylosing Spondylitis Spine Score, have been validated in PsA (98), and a modification of these scales has been proposed for PsA (99).

Portable US technology is increasingly used in clinical practice for both diagnostic purposes (detecting synovitis, joint damage, enthesitis, and dactylitis) as well as for guiding therapeutic injection (100,101). It is being used in an exploratory fashion in clinical trials. US scoring methods for assessing joints, e.g., synovitis and erosions, have been developed for rheumatoid arthritis (RA) and are being explored in PsA (102). In addition, scoring methods for enthesitis are being assessed in PsA (103,104). These efforts are being coordinated through the Outcome Measures in Rheumatology (OMERACT) Ultrasound Working Group.

MRI is increasingly being used as a sensitive imaging technology to detect structural damage and to detect inflammatory change of soft tissue and bone. It is useful both for peripheral pathology (arthritis, enthesitis, and dactylitis) and is particularly useful to assess inflammation and damage in the spine and sacroiliac joints where US is not accurate. An OMERACT MRI working group has developed a scoring system for RA, the Rheumatoid Arthritis Magnetic Resonance Imaging Scoring (RAMRIS) system (105), which is now being adapted for PsA (106). The RAMRIS system was employed and showed responsiveness and discrimination in a trial of abatacept in PsA (107).

COMPOSITE MEASURES OF MULTIPLE DOMAINS IN PSORIATIC ARTHRITIS

Composite response measures, which focus on joint disease as well as patient global/pain, clinician global, and acute-phase reactant (American College of Rheumatology [ACR] response criteria); patient global ± acute-phase reactant (Disease Activity Score [DAS], DAS in 28 joints [DAS28]); and patient/clinician global (Psoriatic Arthritis Response Criteria [PsARC]), have shown reliable discriminant and response characteristics in randomized controlled trials (RCTs), but have not been formally validated in psoriatic arthritis (PsA) (5,12–14). In the past decade, the ACR 20% response criteria (ACR20) have typically been employed as the primary outcome measure of PsA RCTs, and the ACR50 and ACR70, DAS or DAS28, and PsARC have been secondary measures. The ACR and DAS criteria are described in detail in the rheumatoid arthritis section of this article.

Recognizing that PsA is a complex disease that not only involves the domains noted above, but also enthesitis, dactylitis, spine, and skin and nail disease, several groups, including the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis and Outcome Measures in Rheumatology, are working on development of composite measures of disease severity and response to therapy that take into account most, if not all, of these domains. The work on development of these measures is discussed below.

PSORIATIC ARTHRITIS RESPONSE CRITERIA (PSARC)

Description

Purpose. The PsARC was developed as a PsA-specific composite responder index specifically for a study of sulfasalazine in PsA (108). It was first named the PsARC in a subsequent trial of etanercept in PsA (63).

Content. Tender and swollen joint count and patient and physician global assessment.

Number of items. 4 items.

Response options/scale. To achieve response, a patient has to achieve 2 of the following, one of which has to be a tender (68) and swollen (66) joint count, and no worsening of any measure: tender or swollen joint count improvement of ≥30% and/or patient global or physician global improvement of at least 1 point on a 5-point Likert scale.

Recall period for items. At the time of evaluation.

Endorsements. Recommended for use by the European regulatory agency, the European Medicines Agency. However, since there is a tendency for a higher placebo response with this measure and it has lower performance characteristics than the ACR criteria (14), it is not used as the primary outcome measure in PsA RCTs but rather is used as a secondary measure.

Examples of use. It has been used in most PsA clinical trials, as mentioned, as a secondary measure: Mease PJ. Psoriatic arthritis: update on pathophysiology, assessment and management. Ann Rheum Dis 2011;70 Suppl:i77–84 (5).

Mease PJ, Antoni CE. Psoriatic arthritis treatment: biological response modifiers. Ann Rheum Dis 2005;64 Suppl:ii78–82 (8).

Mease PJ. Psoriatic arthritis: pharmacotherapy update. Curr Rheumatol Rep 2010;12:272–80 (9).

Practical Application

How to obtain. Simple scale as described above; no permission necessary.

Method of administration. Physical examination of joints scored on paper, as a Likert scale or patient and physician global.

Scoring. Tender joint count is assessed in 68 joints and swollen joint count in 66. Earlier studies have also employed higher joint counts in PsA, but the current standard is use of the 68/66 joint count. Patient global Likert scale (5 points) is completed in response to the question: "Considering all the ways your arthritis affects you, how are you feeling today?" On the Likert scale, 1 = "very good, no symptoms, and no limitation of normal activity" and 5 = "very poor, very severe symptoms which are intolerable, and inability to carry out normal activities." Physician global assessment involves the same question asked about the patient and assessment on a similarly scored Likert scale.

Score interpretation. Response is achieved if at least 2 of the 4 items are achieved: tender and/or swollen joint count has improved by at least 30% (at least one of these required) and/or patient or physician global has improved by at least 1 point, and no item has worsened.

Respondent burden. Minimal.

Administrative burden. Minimal; takes ~2 minutes if someone to record is present, or slightly longer if not.

Translations/adaptations. A modification of the PsARC using patient and physician global improvement measured by a visual analog scale has been employed and considered acceptable by regulatory agencies (90).

Psychometric Information

Method of development. Derived from expert opinion.

Acceptability. Acceptable.

Reliability. Not assessed.

Validity. The discrimination capability of the PsARC has been evaluated. In a trial of infliximab in PsA, the response rate in the active treatment arm was 82% and in placebo was 30%, yielding a χ^2 value of 27.9 (14). In an etanercept trial, the similar figures were 90% and 33%, respectively, with a χ^2 value of 19.3. The PsARC has not been otherwise formally validated.

Ability to detect change. The ability of individual elements of the PsARC to detect change has been determined. The responsiveness of the quantitative tender and swollen joint count is summarized above. The patient global assessment Likert scale in an etanercept PsA trial demonstrated a standardized response mean (SRM) of −1.51 in the treatment group and 0.48 in placebo, with a t value of 5.8 (14). The physician global assessment Likert scale showed SRMs of −1.98 and −0.34, respectively, for a t value of 7.0 (14).

Critical Appraisal of Overall Value to the Rheumatology Community

Although specifically developed for PsA, the PsARC is limited in that it does not include domains such as enthesitis, dactylitis, or skin disease assessment; tends to display a high placebo response rate; and is primarily used as a secondary measure in clinical trials.

EMERGING COMPOSITE MEASURES

PSORIATIC ARTHRITIS JOINT ACTIVITY INDEX (PSAJAI)

Method of development. The PsAJAI was developed in a project in which data from 3 trials of anti–tumor necrosis factor (anti-TNF) agents in psoriatic arthritis (PsA) were analyzed to create models, based primarily on statistical considerations and some clinical input, which best distinguished active drug from placebo (109,110). Note that in this analysis, addition of the Psoriasis Area and Severity Index (PASI) was problematic in that not all patients in these trials could be assessed for the PASI given low skin scores. Anti-TNF therapy had a large impact on the PASI score; therefore, it was recommended that skin be scored separately. From the same data, response criteria currently used for PsA were examined and logistic regression models based on the individual components of these response criteria were analyzed. The PsAJAI, modeled as the American College of Rheumatology 30% response criteria (ACR30), performed better than the ACR20 and Psoriatic Arthritis Response Criteria (PsARC), and was comparable to previously developed models.

Content. The PsAJAI is the weighted sum of 30% improvement in 6 measures with weights of 2 given to tender joint count, C-reactive protein (CRP) level, and physician global assessment of disease activity. Weights of 1 are given to the remaining 30% improvement measures, including pain, patient global assessment of disease activity, and Health Assessment Questionnaire (110).

Number of items. 6 items.

Recall period for items. Current.

Examples of use. The PsAJAI has not yet been use in a PsA randomized controlled trial (RCT).

Practical Application

How to obtain. Obtainable from the original articles (106,107).

Method of administration. Patient questionnaires and physical examination, recorded on paper or electronically.

Acceptability. Not assessed as of yet.

Reliability. Not determined.

Validity. Not assessed.

Ability to detect change. When applied to a different RCT data set than the 3 from which it was derived, a response rate of 75.6% was noted, similar to a PsARC response of 70% and an ACR20 response of 78% (111).

DISEASE ACTIVITY IN PSORIATIC ARTHRITIS (DAPSA)

Method of development. A Viennese group collected cross-sectional clinical and laboratory data on 105 patients with PsA and performed principal component analysis on those clinical and laboratory variables recommended by the Outcome Measures in Rheumatology (OMERACT) module (112) (Figure 1). Four principal components were derived: patient global and pain visual analog scale (VAS) scores, tender and swollen joint counts, acute-phase reactant (CRP level), and skin, although the latter did not reach statistical significance. The group then studied the existing composite measures and determined that these domains were best served by using the disease activity index for the assessment of reactive arthritis (113). Further, it was determined that the 68/66 joint count outperformed the 28 joint count. This measure was renamed the DAPSA score.

Content. Swollen joint count, tender joint count, patient global, pain, and CRP level.

Number of items. 5 items.

Recall period for items. Current.

Practical Application

How to obtain. Derived from the original article (108).

Method of administration. Patient VAS, physical examination, and laboratory, recorded on paper or electronically.

Scoring. Sum of patient global and pain VAS in centimeters, numerical swollen and tender joint count of 66 and 68 joints, respectively, and CRP level in mg/dl.

Score interpretation. Higher scores reflect more severe disease activity.

Respondent burden. Minimal.

Administrative burden. Takes <5 minutes to obtain.

Psychometric Information

Acceptability. Acceptable.

Reliability. Not yet assessed, although components are commonly used in RCTs and have performed reliably in similar measures.

Validity. The DAPSA was tested in PsA patients starting a new disease-modifying antirheumatic drug treatment (n = 99) and in the data set of a phase III trial of infliximab (Infliximab Multinational Psoriatic Arthritis Controlled Trial 2) (114,115). The instrument correlated highly with other measures, including the Disease Activity Score in 28 joints (DAS28), Simplified Disease Activity Index, and Clinical Disease Activity Index.

Ability to detect change. Effect sizes were high (>0.8) for the active treatment arm and low for placebo in the retrospective analysis of the infliximab data set ($P = 2.56 \times 10^{-10}$), suggesting high discriminant capability (114).

COMPOSITE PSORIATIC DISEASE ACTIVITY INDEX (CPDAI)

Method of development. The CPDAI is based on a grid proposed by the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) to guide treatment decisions in PsA (76,116).

Content. Disease involvement is assessed in up to 5 domains: peripheral joints, skin, enthesial, dactylitis, and spinal manifestations. For each domain, individual instruments are used to assess the extent of disease activity as well as the impact on patient function and health-related quality of life.

How to obtain. Available from the original article (117).

Method of administration. Measures used are patient self-administered, physical examination, and laboratory, recorded on paper or electronically.

Scoring. Domains are scored from 0–3, with empirical cutoffs for disease severity/activity proposed in each largely based on the literature (Table 3). Individual domain scores are summed to give an overall composite score (range 0–15).

Score interpretation. Higher scores correspond to more severe disease activity.

Acceptability. Acceptable.

Reliability. The reliability of the instrument has not been assessed per se. The measures that constitute the scoring system have been routinely used in clinical trials and have been proven to be reliable.

Validity. The CPDAI demonstrates significant correlation with patient (r = 0.777) and physician global (r = 0.809) assessments and discriminates well between effectively and ineffectively treated patients (116).

Ability to detect change. In a cohort of 25 patients in whom treatment was changed, the median CPDAI score had decreased from 8.5 at baseline to 5.5 at 3 months of followup ($P = 0.02$), with a standardized response mean of 0.60.

Critical Appraisal of Overall Value to the Rheumatology Community

The CPDAI has recently been compared to the DAPSA using the Psoriasis Randomized Etanercept Study in Subjects with Psoriatic Arthritis (PRESTA) study data set (118). In PRESTA, 752 patients were randomized to a

Table 3. Modification of the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis grid proposed for the Composite Psoriatic Disease Activity Index (116)*

	Not involved (0)	Mild (1)	Moderate (2)	Severe (3)
Peripheral arthritis		≤4 joints (swollen or tender); normal function (HAQ <0.5)†	≤4 joints but function impaired; or >4 joints, normal function	>4 joints and function impaired
Skin disease		PASI ≤10 and DLQI ≤10	PASI ≤10 but DLQI >10; or PASI >10 but DLQI ≤10	PASI >10 and DLQI >10
Enthesitis		≤3 sites; normal function (HAQ <0.5)†	≤3 sites but function impaired; or >3 sites but normal function	>3 sites and function impaired
Dactylitis		≤3 fingers; normal function (HAQ <0.5)†	≤3 fingers but function impaired; or >3 fingers but normal function	>3 fingers and has function impaired
Spinal disease		BASDAI <4; normal function (ASQOL <6)	BASDAI >4 but normal function; BASDAI <4 but function impaired	BASDAI >4 and function impaired

* HAQ = Health Assessment Questionnaire (58); PASI = Psoriasis Area and Severity Index (19); DLQI = Dermatology Quality of Life Index (71); BASDAI = Bath Ankylosing Spondylitis Disease Activity Index (46); ASQOL = Ankylosing Spondylitis Quality of Life Questionnaire (125).
† HAQ only counted if clinical involvement of domain (joint/enthesis/dactylitis) is present.

double-blind, 2-period study that evaluated the safety and efficacy of 2 doses of etanercept on skin and musculoskeletal disease. Both the CPDAI and DAPSA were effective in determining treatment response in patients treated with etanercept for active psoriasis and PsA. Joint responses were equally determined by both composite scores; however, the CPDAI, which encompasses other domains such as skin, enthesitis, and dactylitis, was the only composite score that could distinguish global treatment response between the 2 etanercept doses. This suggests that the CPDAI is a more sensitive instrument to detect change in domains beyond joints and patient global, particularly in domains such as enthesitis, dactylitis, and the skin, which are important multidimensional components of PsA.

GRACE Project

In addition to critical evaluation of the emerging composite measures described above, the GRAPPA has engaged in a long-term project known as the GRACE Project, in which more than 450 patients with PsA are being evaluated in a multicenter study by members of GRAPPA (Helliwell P: unpublished observations). These patients are being evaluated with multiple measures of all of the clinical domains of PsA longitudinally, capturing information about disease activity and decisions to change therapy in order to determine the individual performance characteristics of measures as well as their ability to provide composite information about PsA as a whole. An attempt is being made to develop a scoring system in which all domains are represented to comprehensively reflect the disease. However, it is acknowledged that this can be a challenge in that musculoskeletal disease activity and response to therapy may be divergent from skin and nail activity and response, and it is critical that the measure be sensitive to disease severity in each domain and not allow "dilution" of a domain in attempting to be comprehensive. It is anticipated that the comparative evaluation of emerging composite measures, including any that are crafted from the GRACE Project, will be presented and voted upon at the OMERACT 12 meeting in May 2012 (Mease PJ: unpublished observations).

Minimal Disease Activity (MDA)

Several studies have utilized remission criteria used in rheumatoid arthritis (RA) to evaluate the ability to achieve this state in PsA. Saber et al comparatively evaluated PsA and RA patients in a single-center study regarding the ability to achieve a DAS28-defined remission (<2.6). They found that a greater percentage of PsA patients than RA patients were able to achieve this degree of response, using a variety of treatment approaches (119). Cantini et al, studying Italian cohorts of PsA and RA patients using the stricter modified ACR criteria for remission from 1981 (120), also found that a greater percentage of PsA patients than RA patients could achieve the state of remission and could remain in that state for a longer period of time (121). These studies suggest that it may be less difficult to aim for sustained remission in PsA than RA. However, these groups have used "joint-centered" definitions of remission, which may be a less comprehensive approach to evaluation of PsA, therefore leading to a GRAPPA project, led by Coates et al, to construct a PsA-specific definition of MDA. Hypothetical cases were evaluated by GRAPPA members and the subsequent analysis by Coates et al resulted in the definition of MDA criteria for PsA shown in Table 4 (122). These criteria were validated by assessing patients in a patient cohort in Toronto (123) and in interventional trial data sets (124). The development of this instrument is a step toward "treatment to target" in PsA, i.e., the goal of achieving remission or low disease activity state.

DISCUSSION

Although relatively young compared to assessment of rheumatoid arthritis and psoriasis, the field of assessment

Table 4. Minimal disease activity (MDA) criteria in psoriatic arthritis (122)*

A patient is classified as in MDA when they meet 5 of 7 of the following criteria:
Tender joint count ≤1
Swollen joint count ≤1
PASI ≤1 or BSA ≤3
Patient pain VAS ≤15
Patient global activity VAS ≤20
HAQ ≤0.5
Tender enthesial points ≤1

* PASI = Psoriasis Area and Severity Index; BSA = body surface area; VAS = visual analog scale; HAQ = Health

of psoriatic arthritis (PsA) has rapidly evolved over the past decade due to the need for valid and reliable assessments in clinical trials of multiple emerging therapeutic agents in PsA as well as growing interest in the disease state. A key factor has been the collaborative endeavors of rheumatologists and dermatologists in the international research consortia, Group for Research and Assessment of Psoriasis and Psoriatic Arthritis, in league with Outcome Measures in Rheumatology. Core domains to be assessed include joint inflammation and damage, enthesitis, dactylitis, skin and nail disease, spondylitis, function, and quality of life. Measures for these individual domains and where available, their performance characteristics, have been described in this article. Additionally, composite measures of disease state and response that have been used in clinical trials and efforts to improve upon these measures that are underway have been described. It is anticipated that as these single-domain and composite measures become codified, simpler and practical measures will evolve for use in clinical practice, allowing for more precise evaluation of disease activity and response to therapy with the goal of achieving remission or minimal disease activity.

ACKNOWLEDGMENT

The author would like to thank Kelly Dundon for assistance in the preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Dr. Mease drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Mease PJ. Improving the routine management of rheumatoid arthritis: the value of tight control. J Rheumatol. 2010;37:1570–8.
2. Schoels M, Knevel R, Aletaha D, Bijlsma JW, Breedveld FC, Boumpas DT, et al. Evidence for treating rheumatoid arthritis to target: results of a systematic literature search. Ann Rheum Dis. 2010;69:638–43.
3. Smolen JS, Aletaha D, Bijlsma JW, Breedveld FC, Boumpas D, Burmester G, et al. Treating rheumatoid arthritis to target: recommendations of an international task force. Ann Rheum Dis 2010;69:631–7.
4. Scott DL, Wolfe F, Huizinga TW. Rheumatoid arthritis. Lancet 2010; 376:1094–108.
5. Mease PJ. Psoriatic arthritis: update on pathophysiology, assessment and management. Ann Rheum Dis 2011;70 Suppl:i77–84.
6. Husni ME, Mease PJ. Managing comorbid disease in patients with psoriatic arthritis. Curr Rheumatol Rep 2010;12:281–7.
7. Gladman DD, Mease PJ, Strand V, Healy P, Helliwell PS, Fitzgerald O, et al. Consensus on a core set of domains for psoriatic arthritis. J Rheumatol 2007;34:1167–70.
8. Mease PJ, Antoni CE. Psoriatic arthritis treatment: biological response modifiers. Ann Rheum Dis 2005;64 Suppl:ii78–82.
9. Mease PJ. Psoriatic arthritis: pharmacotherapy update. Curr Rheumatol Rep 2010;12:272–80.
10. Taylor W, Gladman D, Helliwell P, Marchesoni A, Mease P, Mielants H, and the CASPAR Study Group. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. Arthritis Rheum 2006;54:2665–73.
11. Van Riel P, Van Gestel A, Scott D. EULAR handbook of clinical assessment in rheumatoid arthritis. Alpen an den Rijn (The Netherlands): Van Suiden Communications; 2000.
12. Mease P, Antoni C, Gladman DD, Taylor W. Psoriatic arthritis assessment tools in clinical trials. Ann Rheum Dis 2005;64 Suppl:ii49–54.
13. Gladman DD, Helliwell P, Mease PJ, Nash P, Ritchlin C, Taylor W. Assessment of patients with psoriatic arthritis: a review of currently available measures [review]. Arthritis Rheum 2004;50:24–35.
14. Fransen J, Antoni C, Mease PJ, Uter W, Kavanaugh A, Kalden JR, et al. Performance of response criteria for assessing peripheral arthritis in patients with psoriatic arthritis: analysis of data from randomised controlled trials of two tumour necrosis factor inhibitors. Ann Rheum Dis 2006;65:1373–8.
15. Englbrecht M, Wang Y, Ronneberger M, Manger B, Vastesaeger N, Veale DJ, et al. Measuring joint involvement in polyarticular psoriatic arthritis: an introduction of alternatives. Arthritis Care Res (Hoboken) 2010;62:977–83.
16. Gladman DD, Inman RD, Cook RJ, Maksymowych WP, Braun J, Davis JC, et al. International spondyloarthritis interobserver reliability exercise: the INSPIRE study. II. Assessment of peripheral joints, enthesitis, and dactylitis. J Rheumatol 2007;34:1740–5.
17. Chandran V, Gottlieb A, Cook RJ, Duffin KC, Garg A, Helliwell P, et al. International multicenter psoriasis and psoriatic arthritis reliability trial for the assessment of skin, joints, nails, and dactylitis. Arthritis Rheum 2009;61:1235–42.
18. Buskila D, Langevitz P, Gladman DD, Urowitz S, Smythe HA. Patients with rheumatoid arthritis are more tender than those with psoriatic arthritis. J Rheumatol 1992;19:1115–9.
19. Fredriksson T, Pettersson U. Severe psoriasis: oral therapy with a new retinoid. Dermatologica 1978;157:238–44.
20. Long CC, Finlay AY. The finger-tip unit: a new practical measure. Clin Exp Dermatol 1991;16:444–7.
21. Long CC, Finlay AY, Averill RW. The rule of hand: 4 hand areas = 2 FTU = 1 g. Arch Dermatol 1992;128:1129–30.
22. Langley RG, Ellis CN. Evaluating psoriasis with psoriasis area and severity index, psoriasis global assessment, and lattice system physician's global assessment. J Am Acad Dermatol 2004;51:563–9.
23. Krueger GG. New method being developed for assessing psoriasis. National Psoriasis Foundation Forum 1999;5:4–5.
24. Berth-Jones J, Thompson J, Papp K. A study examining inter-rater and intrarater reliability of a novel instrument for assessment of psoriasis: the Copenhagen Psoriasis Severity Index. Br J Dermatol 2008;159:407–12.
25. Ashcroft DM, Wan Po AL, Williams HC, Griffiths CE. Clinical measures of disease severity and outcome in psoriasis: a critical appraisal of their quality. Br J Dermatol 1999;141:185–91.
26. Louden BA, Pearce DJ, Lang W, Feldman SR. A simplified psoriasis area severity index (SPASI) for rating psoriasis severity in clinic patients. Dermatol Online J 2004;10:7.
27. Feldman SR, Fleischer AB Jr, Reboussin DM, Rapp SR, Exum ML, Clark AR, et al. The self-administered Psoriasis Area and Severity Index is valid and reliable. J Invest Dermatol 1996;106:183–6
28. Carlin CS, Feldman SR, Krueger JG, Menter A, Krueger GG. A 50% reduction in the Psoriasis Area and Severity Index (PASI 50) is a clinically significant endpoint in the assessment of psoriasis. J Am Acad Dermatol 2004;50:859–66.
29. Rich P, Scher RK. Nail psoriasis severity index: a useful tool for evaluation of nail psoriasis. J Am Acad Dermatol 2003;49:206–12.
30. Aktan S, Ilknur T, Akin C, Ozkan S. Interobserver reliability of the Nail Psoriasis Severity Index. Clin Exp Dermatol 2007;32:141–4.
31. Cassell SE, Bieber JD, Rich P, Tutuncu ZN, Lee SJ, Kalunian KC, et al. The modified nail psoriasis severity index: validation of an instrument to assess psoriatic nail involvement in patients with psoriatic arthritis. J Rheumatol 2007;34:123–9.
32. Augustin M, Ogilvie A. Methods of outcomes measurement in nail psoriasis. Dermatology 2010;221 Suppl:23–8.
33. Kavanaugh A, McInnes I, Mease P, Krueger GG, Gladman D, Gomez-Reino J, et al. Golimumab, a new human tumor necrosis factor α antibody, administered every four weeks as a subcutaneous injection in psoriatic arthritis: twenty-four–week efficacy and safety results of a

randomized, placebo-controlled study. Arthritis Rheum 2009;60:976–86.

34. McGonagle D, Lories RJ, Tan AL, Benjamin M. The concept of a "synovio-entheseal complex" and its implications for understanding joint inflammation and damage in psoriatic arthritis and beyond. Arthritis Rheum 2007;56:2482–91.
35. Gladman DD, Chandran V. Observational cohort studies: lessons learnt from the University of Toronto Psoriatic Arthritis Program. Rheumatology (Oxford) 2011;50:25–31.
36. Mander M, Simpson JM, McLellan A, Walker D, Goodacre JA, Dick WC. Studies with an enthesis index as a method of clinical assessment in ankylosing spondylitis. Ann Rheum Dis 1987;46:197–202.
37. Healy PJ, Helliwell PS. Measuring clinical enthesitis in psoriatic arthritis: assessment of existing measures and development of an instrument specific to psoriatic arthritis. Arthritis Rheum 2008;59: 686–91.
38. Maksymowych WP, Mallon C, Morrow S, Shojania K, Olszynski WP, Wong RL, et al. Development and validation of the Spondyloarthritis Research Consortium of Canada (SPARCC) Enthesitis Index. Ann Rheum Dis 2009;68:948–53.
39. Heuft-Dorenbosch L, Spoorenberg A, van Tubergen A, Landewe R, van ver Tempel H, Mielants H, et al. Assessment of enthesitis in ankylosing spondylitis. Ann Rheum Dis 2003;62:127–32.
40. Braun J, Brandt J, Listing J, Zink A, Alten R, Golder W, et al. Treatment of active ankylosing spondylitis with infliximab: a randomised controlled multicentre trial. Lancet 2002;359:1187–93.
41. Gorman JD, Sack KE, Davis JC Jr. Treatment of ankylosing spondylitis by inhibition of tumor necrosis factor α. N Engl J Med 2002;346:1349–56.
42. Coates LC, Helliwell PS. Disease measurement: enthesitis, skin, nails, spine and dactylitis. Best Pract Res Clin Rheumatol 2010;24:659–70.
43. Helliwell PS, Firth J, Ibrahim GH, Melsom RD, Shah I, Turner DE. Development of an assessment tool for dactylitis in patients with psoriatic arthritis. J Rheumatol 2005;32:1745–50.
44. Furst DE, Keystone EC, Braun J, Breedveld FC, Burmester GR, De Benedetti F, et al. Updated consensus statement on biological agents for the treatment of rheumatic diseases, 2010. Ann Rheum Dis 2011;70 Suppl:i2–36.
45. Healy PJ, Helliwell PS. Measuring dactylitis in clinical trials: which is the best instrument to use? J Rheumatol 2007;34:1302–6.
46. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J Rheumatol 1994;21:2286–91.
47. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994;21:2281–5.
48. Jenkinson TR, Mallorie PA, Whitelock HC, Kennedy LG, Garrett SL, Calin A. Defining spinal mobility in ankylosing spondylitis (AS): the Bath AS Metrology Index. J Rheumatol 1994;21:1694–8.
49. Taylor WJ, Harrison AA. Could the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) be a valid measure of disease activity in patients with psoriatic arthritis? Arthritis Rheum 2004;51: 311–5.
50. Leung YY, Tam LS, Kun EW, Ho KW, Li EK. Comparison of 4 functional indexes in psoriatic arthritis with axial or peripheral disease subgroups using Rasch analyses. J Rheumatol 2008;35:1613–21.
51. Gladman D, Inman R, Cook RJ, van der Heijde D, Landewe RB, Braun J, et al. International spondyloarthritis interobserver reliability exercise: the INSPIRE study. I. Assessment of spinal measures. J Rheumatol 2007;34:1733–9.
52. Fernandez-Sueiro JL, Willisch A, Pertega-Diaz S, Tasende JA, Fernandez-Lopez C, Galdo F, et al. Evaluation of ankylosing spondylitis spinal mobility measurements in the assessment of spinal involvement in psoriatic arthritis. Arthritis Rheum 2009;61:386–92.
53. Lukas C, Landewe R, Sieper J, Dougados M, Davis J, Braun J, et al. Development of an ASAS-endorsed disease activity score (ASDAS) in patients with ankylosing spondylitis. Ann Rheum Dis 2009;68:18–24.
54. Eder L, Chandran V, Shen H, Cook RJ, Gladman DD. Is ASDAS better than BASDAI as a measure of disease activity in axial psoriatic arthritis? Ann Rheum Dis 2010;69:2160–4.
55. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.
56. Cauli A, Gladman DD, Mathieu A, Olivieri I, Porru G, Tak PP, et al. Patient global assessment in psoriatic arthritis: a multicenter GRAPPA and OMERACT study. J Rheumatol 2011;38:898–903.
57. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 2008; 9:105–21.
58. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
59. Blackmore MG, Gladman DD, Husted J, Long JA, Farewell VT. Measuring health status in psoriatic arthritis: the Health Assessment Questionnaire and its modification. J Rheumatol 1995;22:886–93.
60. Husted JA, Gladman DD, Long JA, Farewell VT. A modified version of the Health Assessment Questionnaire (HAQ) for psoriatic arthritis. Clin Exp Dermatol 1995;13:439–43.
61. Husted JA, Tom BD, Farewell VT, Schentag CT, Gladman DD. A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis: does the effect change over time? Arthritis Rheum 2007;56:840–9.
62. Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. Arthritis Rheum 2007;57:723–9.
63. Mease PJ, Goffe BS, Metz J, VanderStoep A, Finck B, Burge DJ. Etanercept in the treatment of psoriatic arthritis and psoriasis: a randomised trial. Lancet 2000;356:385–90.
64. Mease PJ, Woolley JM, Bitman B, Wang B, Globe D, Singh A. Minimally important difference of Health Assessment Questionnaire in psoriatic arthritis: relating thresholds of improvement in functional ability to patient-rated importance and satisfaction. J Rheumatol. E-pub ahead of print.
65. Mease P, Ganguly L, Wanke E, Yu E, Singh A. How much improvement in functional status is considered important by patients with active psoriatic arthritis: applying the Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) group guidelines [abstract]. Ann Rheum Dis 2004;63 Suppl:391.
66. Kwok T, Pope JE. Minimally important difference for patient-reported outcomes in psoriatic arthritis: Health Assessment Questionnaire and pain, fatigue, and global visual analog scales. J Rheumatol 2010;37: 1024–8.
67. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE Jr. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum 2000;43:1478–87.
68. Husted J, Gladman DD, Long JA, Farewell VT. Relationship of the Arthritis Impact Measurement Scales to changes in articular status and functional performance in patients with psoriatic arthritis. J Rheumatol 1996;23:1932–7.
69. Husted J, Gladman DD, Farewell VT, Long JA. Validation of the revised and expanded version of the Arthritis Impact Measurement Scales for patients with psoriatic arthritis. J Rheumatol 1996;23: 1015–9.
70. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
71. Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI): a simple practical measure for routine clinical use. Clin Exp Dermatol 1994;19: 210–6.
72. The EuroQol Group. EuroQol: a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
73. Husted JA, Gladman DD, Farewell VT, Long JA, Cook RJ. Validating the SF-36 health survey questionnaire in patients with psoriatic arthritis. J Rheumatol 1997;24:511–7.
74. Nichol MB, Margolies JE, Lippa E, Rowe M, Quell J. The application of multiple quality-of-life instruments in individuals with mild-to-moderate psoriasis. Pharmacoeconomics 1996;10:644–53.
75. Kurwa HA, Finlay AY. Dermatology in-patient management greatly improves life quality. Br J Dermatol 1995;133:575–8.
76. Ritchlin CT, Kavanaugh A, Gladman DD, Mease PJ, Helliwell P, Boehncke WH, et al. Treatment recommendations for psoriatic arthritis. Ann Rheum Dis 2009;68:1387–94.
77. McKenna SP, Doward LC, Whalley D, Tennant A, Emery P, Veale DJ. Development of the PsAQoL: a quality of life instrument specific to psoriatic arthritis. Ann Rheum Dis 2004;63:162–9.
78. Healy PJ, Helliwell PS. Psoriatic arthritis quality of life instrument: an assessment of sensitivity and response to change. J Rheumatol 2008; 35:1359–61.
79. Stamm TA, Nell V, Mathis M, Coenen M, Aletaha D, Cieza A, et al. Concepts important to patients with psoriatic arthritis are not adequately covered by standard measures of functioning. Arthritis Rheum 2007;57:487–94.
80. Kirwan JR, Newman S, Tugwell PS, Wells GA. Patient perspective on outcomes in rheumatology: a position paper for OMERACT 9. J Rheumatol 2009;36:2067–70.
81. Minnock P, Kirwan J, Veale D, Fitzgerald O, Bresnihan B. Fatigue is an independent outcome measure and is sensitive to change in patients with psoriatic arthritis. Clin Exp Rheumatol 2010;28:401–4.
82. Belza BL, Henke CJ, Yelin EH, Epstein WV, Gilliss CL. Correlates of

fatigue in older adults with rheumatoid arthritis. Nurs Res 1993;42: 93–9.

83. Smets EM, Garssen B, Bonke B, De Haes JC. The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39:315–25.
84. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The Fatigue Severity Scale: application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol 1989;46:1121–3.
85. Cella D, Lai JS, Chang CH, Peterman A, Slavin M. Fatigue in cancer patients compared with fatigue in the general United States population. Cancer 2002;94:528–38.
86. Mendoza TR, Wang XS, Cleeland CS, Morrissey M, Johnson BA, Wendt JK, et al. The rapid assessment of fatigue severity in cancer patients: use of the brief fatigue inventory. Cancer 1999;85:1186–96.
87. Wolfe F, Hawley DJ, Wilson K. The prevalence and meaning of fatigue in rheumatic disease. J Rheumatol 1996;23:1407–17.
88. Schentag CT, Cichon J, MacKinnon A, Gladman DD, Urowitz MB. Validation and normative data for the 0-10 point scale version of the fatigue severity scale (FSS) [abstract]. Arthritis Rheum 2000;43 Suppl: S177.
89. Schentag C, Gladman D. Changes in fatigue in psoriatic arthritis: disease activity of fibromyalgia [abstract]. Arthritis Rheum 2002;46 Suppl:S424.
90. Mease PJ, Gladman DD, Ritchlin CT, Ruderman EM, Steinfeld SD, Choy EH, et al, for the Adalimumab Effectiveness in Psoriatic Arthritis Trial Study Group. Adalimumab for the treatment of patients with moderately to severely active psoriatic arthritis: results of a double-blind, randomized, placebo-controlled trial. Arthritis Rheum 2005;52: 3279–89.
91. Chandran V, Bhella S, Schentag C, Gladman DD. Functional assessment of chronic illness therapy-fatigue scale is valid in patients with psoriatic arthritis. Ann Rheum Dis 2007;66:936–9.
92. Cella D. Manual of the Functional Assessment of Chronic Illness Therapy (FACIT) measurement system. Evanston (IL): Center on Outcomes, Research, and Education, Evanston Northwestern Healthcare and Northwestern University; 1997.
93. Webster K, Cella D, Yost K. The Functional Assessment of Chronic Illness Therapy (FACIT) measurement system: properties, applications, and interpretation. Health Qual Life Outcomes 2003;1:79.
94. Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E. Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. J Pain Symptom Manage 1997;13:63–74.
95. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45 Suppl:S22–31.
96. Mease P, van der Heijde D. Joint damage in psoriatic arthritis: how is it assessed and can it be prevented? Int J Adv Rheumatol 2006;4:38–48.
97. Van der Heijde D, Sharp J, Wassenberg S, Gladman DD. Psoriatic arthritis imaging: a review of scoring methods. Ann Rheum Dis 2005;64 Suppl:ii61–4.
98. Lubrano E, Marchesoni A, Olivieri I, D'Angelo S, Spadaro A, Parsons WJ, et al. The radiological assessment of axial involvement in psoriatic arthritis: a validation study of the BASRI total and the modified SASSS scoring methods. Clin Exp Rheumatol 2009;27:977–80.
99. Lubrano E, Marchesoni A, Olivieri I, D'Angelo S, Spadaro A, Parsons WJ, et al. Psoriatic Arthritis Spondylitis Radiology Index: a modified index for radiologic assessment of axial involvement in psoriatic arthritis. J Rheumatol 2009;36:1006–11.
100. Sturrock RD. Clinical utility of ultrasonography in spondyloarthropathies. Curr Rheumatol Rep 2009;11:317–20.
101. Kane D. The role of ultrasound in the diagnosis and management of psoriatic arthritis. Curr Rheumatol Rep 2005;7:319–24.
102. Dougados M, Jousse-Joulin S, Mistretta F, d'Agostino MA, Backhaus M, Bentin J, et al. Evaluation of several ultrasonography scoring systems for synovitis and comparison to clinical examination: results from a prospective multicentre study of rheumatoid arthritis. Ann Rheum Dis 2010;69:828–33.
103. Balint PV, Kane D, Wilson H, McInnes IB, Sturrock RD. Ultrasonography of entheseal insertions in the lower limb in spondyloarthropathy. Ann Rheum Dis 2002;61:905–10.
104. Gisondi P, Tinazzi I, El-Dalati G, Gallo M, Biasi D, Barbara LM, et al. Lower limb enthesopathy in patients with psoriasis without clinical signs of arthropathy: a hospital-based case-control study. Ann Rheum Dis 2008;67:26–30.
105. Ostergaard M, Peterfy C, Conaghan P, McQueen F, Bird P, Ejbjerg B, et al. OMERACT rheumatoid arthritis magnetic resonance imaging studies: core set of MRI acquisitions, joint pathology definitions, and the OMERACT RA-MRI scoring system [published erratum appears in J Rheumatol 2004;31:198]. J Rheumatol 2003;30:1385–6.
106. McQueen F, Lassere M, Bird P, Haavardsholm EA, Peterfy C, Conaghan PG, et al. Developing a magnetic resonance imaging scoring system for peripheral psoriatic arthritis. J Rheumatol 2007;34:859–61.
107. Mease P, Genovese MC, Gladstein G, Kivitz AJ, Ritchlin C, Tak PP, et al. Abatacept in the treatment of patients with psoriatic arthritis: results of a six-month, multicenter, randomized, double-blind, placebo-controlled, phase II trial. Arthritis Rheum 2011;63:939–48.
108. Clegg DO, Reda DJ, Mejias E, Cannon GW, Weisman MH, Taylor T, et al. Comparison of sulfasalazine and placebo in the treatment of psoriatic arthritis: a Department of Veterans Affairs cooperative study. Arthritis Rheum 1996;39:2013–20.
109. Gladman DD, Tom BD, Mease PJ, Farewell VT. Informing response criteria for psoriatic arthritis. I. Discrimination models based on data from 3 anti-tumor necrosis factor randomized studies. J Rheumatol 2010;37:1892–7.
110. Gladman DD, Tom BD, Mease PJ, Farewell VT. Informing response criteria for psoriatic arthritis (PsA). II. Further considerations and a proposal: the PsA joint activity index. J Rheumatol 2010;37:2559–65.
111. Gladman DD, Psaradellis F, Illouz O, Sampalis JS. Evaluation of response using the psoriatic arthritis joint activity index scoring tool in patients treated with adalimumab: post hoc analysis of the AC-CLAIM study [abstract]. J Rheumatol 2009;38:2571.
112. Nell-Duxneuner VP, Stamm TA, Machold KP, Pflugbeil S, Aletaha D, Smolen JS. Evaluation of the appropriateness of composite disease activity measures for assessment of psoriatic arthritis. Ann Rheum Dis 2010;69:546–9.
113. Eberl G, Studnicka-Benke A, Hitzelhammer H, Gschnait F, Smolen JS. Development of a disease activity index for the assessment of reactive arthritis (DAREA). Rheumatology (Oxford) 2000;39:148–55.
114. Schoels M, Aletaha D, Funovits J, Kavanaugh A, Baker D, Smolen JS. Application of the DAREA/DAPSA score for assessment of disease activity in psoriatic arthritis. Ann Rheum Dis 2010;69:1441–7.
115. Antoni C, Krueger GG, de Vlam K, Birbara C, Beutler A, Guzzo C, et al. Infliximab improves signs and symptoms of psoriatic arthritis: results of the IMPACT 2 trial. Ann Rheum Dis 2005;64:1150–7.
116. Mumtaz A, Gallagher P, Kirby B, Waxman R, Coates LC, Veale JD, et al. Development of a preliminary composite disease activity index in psoriatic arthritis. Ann Rheum Dis 2010;70:272–7.
117. Mumtaz A, Gallagher P, Kirby B, Waxman R, Coates LC, Veale JD, et al. Development of a preliminary composite disease activity index in psoriatic arthritis. Ann Rheum Dis 2010;70:272–7.
118. FitzGerald O, Helliwell P, Mumtaz A, Coates L, Pedersen R, Molta C. Application of composite disease activity scores in psoriatic arthritis to the PRESTA dataset [abstract]. Arthritis Rheum 2010;62 Suppl: S214.
119. Saber TP, Ng CT, Renard G, Lynch BM, Pontifex E, Walsh CA, et al. Remission in psoriatic arthritis: is it possible and how can it be predicted? Arthritis Res Ther 2010;12:R94.
120. Pinals RS, Masi AT, Larsen RA, and the Subcommittee for Criteria of Remission in Rheumatoid Arthritis of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Preliminary criteria for clinical remission in rheumatoid arthritis. Arthritis Rheum 1981;24:1308–15.
121. Cantini F, Niccoli L, Nannini C, Cassara E, Pasquetti P, Olivieri I, et al. Frequency and duration of clinical remission in patients with peripheral psoriatic arthritis requiring second-line drugs. Rheumatology (Oxford) 2008;47:872–6.
122. Coates LC, Fransen J, Helliwell PS. Defining minimal disease activity in psoriatic arthritis: a proposed objective target for treatment. Ann Rheum Dis 2010;69:48–53.
123. Coates LC, Cook R, Lee KA, Chandran V, Gladman DD. Frequency, predictors, and prognosis of sustained minimal disease activity in an observational psoriatic arthritis cohort. Arthritis Care Res (Hoboken) 2010;62:970–6.
124. Coates LC, Helliwell PS. Validation of minimal disease activity criteria for psoriatic arthritis using interventional trial data. Arthritis Care Res (Hoboken) 2010;62:965–9.
125. Doward LC, Spoorenberg A, Cook SA, Whalley D, Helliwell PS, Kay LJ, et al. Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. Ann Rheum Dis 2003;62:20–6.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S200–S207
DOI 10.1002/acr.20549

Measures of Hip Function and Symptoms

Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity for Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire

ANNA NILSDOTTER[1] AND ANN BREMANDER[2]

INTRODUCTION

Outcome measures included in this review are the Harris Hip Score, the Hip Disability and Osteoarthritis Outcome Score, the Oxford Hip Score, the Lequesne Index of Severity for Osteoarthritis of the Hip, and the American Academy of Orthopedic Surgeons Hip and Knee Questionnaire.

The outcome measures chosen are the most common ones in the literature concerning hip function and symptoms. Most of them are patient-reported. The selected measures meet the basic requirements for an outcome measurement, although there are shortcomings in a few of them.

HARRIS HIP SCORE (HHS)

Description

Purpose. The HHS was developed for the assessment of the results of hip surgery, and is intended to evaluate various hip disabilities and methods of treatment (1) in an adult population. The original version was published 1969.

Content. The domains covered are pain, function, absence of deformity, and range of motion. The pain domain measures pain severity and its effect on activities and need for pain medication.

The function domain consists of daily activities (stair use, using public transportation, sitting, and managing shoes and socks) and gait (limp, support needed, and walking distance). Deformity takes into account hip flexion, adduction, internal rotation, and extremity length discrepancy. Range of motion measures hip flexion, abduction, external and internal rotation, and adduction.

Number of items. There are 10 items.

Response options/scale. The score has a maximum of 100 points (best possible outcome) covering pain (1 item, 0–44 points), function (7 items, 0–47 points), absence of deformity (1 item, 4 points), and range of motion (2 items, 5 points).

Recall period for items. Not described.

Examples of use. Total hip replacement (THR) (1–4), femoral neck fractures (5), and osteoarthritis (6).

Practical Application

How to obtain. Available in the original article (1), URL: http://www.orthopaedicscore.com/ and URL: http://www.ncbi.nlm.nih.gov/pubmed/5783851.

Method of administration. The HHS is a clinician-based outcome measure administered by a qualified health care professional, such as a physician or a physical therapist.

Scoring. Each item has a unique numerical scale, which corresponds to descriptive response options. The number of response options varies by item, as does the number of points assigned to each response option. The range of motion item consists of 6 motions that are graded based on the arc of motion possible. Each range of motion gradation is assigned an index factor and a maximum possible value, which are used to calculate arc of motion points. These points are added and multiplied by 0.05 to receive the total points for range of motion. The total score is calculated by summing the scores for the 4 domains.

Score interpretation. The HHS score gives a maximum of 100 points. Pain receives 44 points, function 47 points, range of motion 5 points, and deformity 4 points. Function is subdivided into activities of daily living (14 points) and gait (33 points).

The higher the HHS, the less dysfunction. A total score of <70 is considered a poor result; 70–80 is considered fair, 80–90 is good, and 90–100 is an excellent result (1). No normative values are available.

Respondent burden. Takes 5 minutes to complete.

[1]Anna Nilsdotter, PhD: Halmstad Central Hospital, Halmstad, Sweden; [2]Ann Bremander, PT, PhD: Lund University, Lund, and Spenshult Hospital for Rheumatic Diseases, Oskarström, Sweden.

Address correspondence to Ann Bremander, PT, PhD, Research and Development Center, Spenshult Hospital for Rheumatic Diseases, SE 313 92 Oskarström, Sweden. E-mail: ann.bremander@spenshult.se.

Submitted for publication March 4, 2011; accepted in revised form June 30, 2011.

Administrative burden. No formal training is necessary. Data calculating can be performed automatically during data processing using computer-based algorithms.

Translations/adaptations. The HHS has been used in many different countries (Sweden, The Netherlands, Denmark, etc.), but there are no validated versions in other languages available.

Psychometric Information

Method of development. Thirty-eight (31 men) individuals who had undergone THR operations due to traumatic arthritis were the first patients who were evaluated with the HHS. The items were generated based on the opinion by experts that pain and functional capacity are the 2 basic considerations. They were the indications for surgery and hence received the heaviest weighting: 91 of 100 points (1).

Acceptability. Wamper et al (7) report unacceptable ceiling effects in 31 of 59 studies. Pooled data across the studies included (n = 6,667 patients) suggested ceiling effects of 20% (95% confidence interval 18–22).

Reliability. Cronbach's alpha coefficient showed high internal consistency reliability except for deformity, which could not be calculated.

The test–retest interval was 3 to 4 weeks. The total score reliability was excellent for physicians (r = 0.94) and physiotherapists (r = 0.95). The physiotherapist and the orthopedic surgeon showed excellent test–retest reliability in the domains of pain (r = 0.93 and r = 0.98, respectively) and function (r = 0.95 and r = 0.93, respectively). The calculations were done with Pearson's and Spearman's correlation coefficients (8).

The interrater correlations were good to excellent (0.74–1.0) for the domain scores in Söderman's study, as well as in study by Kirmit et al (8,9).

Validity. The HHS content validity has been tested by directly comparing HHS, the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), and the Short Form 36 (SF-36). No major differences between the scores were seen (8). The HHS construct validity was tested by comparing the pain and function domains in HHS, WOMAC, Nottingham Health Profile (NHP), and SF-36. The HHS domains pain and function correlated (Spearman's rho) better with similar domains in WOMAC, NHP, and SF-36 than with different domains (4). In another study, the same result was obtained when comparing HHS, WOMAC, and SF-36 (8). Correlations (Kendall's tau) between HHS and SF-36 have been shown to be strong in the physical domains (3) and weak in the mental domains. A strong correlation (Spearman's rho) has been found between HHS and NHP (2).

Ability to detect change. HHS responsiveness has been determined in a study of 335 THRs. The effect size between preoperative and 6-months postoperative was excellent for pain (2.80) and function (1.72), but weak in the 2-years followup, i.e., pain (0.15) and function (0.18) (10). When comparing the HHS, Barthel Index, and EuroQol 5-domain (EQ-5D) in patients with femoral neck fractures 4 and 12 months after surgery, the standardized response mean was 0.75 for HHS, 0.40 for Barthel Index, and 0.46 for EQ-5D (5).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The HHS is widely used throughout the world for evaluating outcome after THR (11). The indication for THR is particularly pain and impaired physical function, which are the 2 dominating domains in HHS. The HHS has also been proven appropriate to measure outcome after interventions such as physical therapy (6) and femoral neck fractures (5).

Caveats and cautions. There are unacceptable ceiling effects that severely limit its validity (7).

Clinical usability. The psychometric evaluation does not support interpretation of scores to make decisions for individuals. The administrative burden does not limit clinical use nor the respondent burden since HHS is not self-administered.

Research usability. For short-time followup studies it seems to be useful (5,10) if you are aware of the problem with the ceiling effects.

HIP DISABILITY AND OSTEOARTHRITIS OUTCOME SCORE (HOOS)

Description

Purpose. HOOS was developed as an instrument to assess the patients' opinion about their hip and associated problems, and is intended to be used in an adult population with hip disability with or without osteoarthritis (OA).

HOOS has been validated in 2 slightly different versions, LK 1.1 and LK 2.0 (12,13). The LK 2.0 version is available on line at www.koos.nu. HOOS includes Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) LK 3.0 (14) in its complete and original format (with permission), and WOMAC scores can be calculated. In 2008, a 5-item measure of physical function, the HOOS-PS, was published derived from the HOOS questionnaire by item-response theory to elicit patients' opinions about difficulties experienced due to hip problems (15).

Content. HOOS consists of 5 subscales: pain, other symptoms, function in activities of daily living (ADL), and function in sport and recreation (Sport/Rec), and hip-related quality of life (QOL).

Number of items. In total, 40 items: 10 items for pain, 5 items for other symptoms (3 for symptoms and 2 for stiffness), 17 items for function in ADL, 4 items for function in Sport/Rec, and 4 items for hip-related QOL.

Response options/scale. Standardized answer options are given (5 Likert boxes) and each question is scored from 0 to 4. Scores are summarized for each subscale and transformed to a 0–100 scale (0 indicating extreme problems and 100 indicating no problems).

Recall period for items. The last week is taken into consideration when answering the questions.

Endorsements. The HOOS-PS was the result of an Osteoarthritis Research Society International (OARSI) and

Outcome Measures in Rheumatology (OMERACT) initiative (15).

Examples of use. The HOOS has been used in subjects with hip disability with or without hip osteoarthritis (12), and in patients with hip OA pre- and postoperative total hip replacement (THR) (13,15).

Practical Application

How to obtain. The HOOS can be obtained for no cost at: www.koos.nu.

Method of administration. The questionnaire is patient reported.

Scoring. The user's guide includes a manual scoring sheet and an Excel file ready to download at the web site (www.koos.nu). There are instructions for handling missing values in the user's guide. Computer scoring is not necessary but is recommended since it increases the usefulness in the clinic.

Score interpretation. Each subscale has a score of 0–100, where 0 indicates extreme problems and 100 indicates no problems. The results can be plotted as an outcome profile, the HOOS profile. (The HOOS-PS was used in an OARSI-OMERACT–supported study of pain and functional disability and its correspondence to total joint replacement. Neither pain nor functional disability alone could discriminate between patients who were or were not eligible for a total joint replacement according to the orthopedic surgeon [16].)

Respondent burden. The HOOS questionnaire takes ~10–15 minutes to complete.

Administrative burden. No administration burden; time to score by hand takes 10–15 minutes. No training is necessary. Computer scoring by using the Excel file only takes 2 or 3 minutes (entering of data).

Translations/adaptations. Available in Swedish (13), Dutch (17), and French (18) all with published validation studies. Available in Danish, English, German, Korean, and Lithuanian, according to the web site. Also available in versions for knee injury and knee OA (Knee injury and Osteoarthritis Outcome Score), for a variety of foot- and ankle-related problems (Foot and Ankle Outcome Score), and for assessing problems from the lower extremity in patients with inflammatory arthritis (Rheumatoid and Arthritis Outcome Score). All information is available at the web site, www.koos.nu.

Psychometric Information

Method of development. Items in the HOOS questionnaire were generated through literature search, through interviews with more than 100 patients with hip disability, with and without hip OA (12), and by questioning 90 patients undergoing THR (13).

Acceptability. Missing data are reported to range from 0.9–2.6% in the different validations studies (13,18). A total score could be calculated for 99% of the subjects in the Swedish validation study and for all subjects in the French study.

Floor effects are more common in the subscale Sport/Rec, where worst possible scores have been reported to range from 4.1–17.8% in subjects eligible for THR and subjects with hip OA (13,17,18). Reports of ceiling effects have only been reported in the Swedish validation study 6 months after THR where 19% of the subjects reported a best possible score in the pain subscale, 10% in the symptoms subscale, 5% in the ADL subscale, and 9% in the Sport/Rec and the QOL subscale (13).

Reliability. HOOS has been used in patients ages 42–89 years, including subjects with hip OA treated by medication only, subjects eligible for THR and postoperatively (12,13,17,18). The internal consistency ranged from 0.82 to 0.98 (Cronbach's alpha coefficient) in the different studies (12,17,18), with the highest value in the ADL subscale (0.94–0.98), which might indicate a redundancy of items. HOOS has high test–retest reproducibility, with the intraclass correlation coefficient ranging from 0.75 to 0.97 in the validation studies (12,17,18).

The standard error of measurement published in the Dutch study ranged from 3.71 (QOL subscale) to 6.94 (pain subscale) for subjects with hip OA, and from 4.78 (ADL subscale) to 10.07 (Sport/Rec subscale) for subjects who had undergone THR (17).

Validity. HOOS content validity was performed by asking patients to rate item importance in the 2 Swedish validation studies (12,13) resulting in slightly different questionnaires where the LK 2.0 version has been translated into Dutch and French. HOOS construct validity has been tested by comparing it with the Short Form 36, the Oxford Hip Score, the Lequesne Index, and the visual analog scale for pain, and predetermined hypotheses were confirmed (13,17,18).

Ability to detect change. HOOS responsiveness has been determined in 1 Swedish and in 1 French study (n = 90 and n = 30, respectively) after THR (13,18). The standardized response mean ranged from 1.29–3.24 (13,18). Younger patients (age <66 years) showed larger responsiveness in all subscales compared with older subjects (13).

In the French sample, the effect size ranged from 1.97 (QOL subscale) to 3.24 (pain subscale) (18). The smallest detectable difference of the HOOS ranged from 9.6 for the ADL subscale to 16.2 for the QOL subscale (18).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The HOOS is an extension of the WOMAC and is suggested to be valuable for younger and more active people due to added subscales. The HOOS has been included in 2 systematic reviews concerning psychometric evaluations of questionnaires assessing hip OA and yielded positive findings (19,20). The HOOS needs further psychometric testing in different cultures and in different groups of patients with hip disabilities.

Clinical usability. The HOOS can be used to follow patients with hip OA over time in the clinic, whatever the severity. Using the Excel file at the web site to calculate scores makes it fast and easy to administer.

Research usability. HOOS is suitable to use in research as a disease-specific questionnaire.

OXFORD HIP SCORE (OHS)

Description

Purpose. To assess outcome after total hip replacement (THR) by measuring patients' perceptions in adjunction to surgery. The original version from 1996 (21) was updated in 2007 introducing a new scoring system (22).

Content. OHS assesses pain (6 items) and function (6 items) of the hip in relation to daily activities such as walking, dressing, sleeping, etc.

Number of items. 12 items with 5 categories of response; no subscales.

Response options/scale. The original scoring from 1996 ranged from 1–5 (best to worst) with a total score of 12–60 (least difficulties to most difficulties) (21).

A new scoring was suggested in 2007 and supported by the original authors: 0–4 (worst to best) with overall scores ranging from 0–48 where 48 represents the best score (22).

Recall period for items. During the past 4 weeks.

Examples of use. Designed for assessment of joint replacement and has been used in several countries in large registry studies (23–28). Has also been validated and used in revision hip replacement (29,30).

Practical Application

How to obtain. Information concerning the Oxford Orthopaedic scores can be found at http://phi.uhce.ox.ac.uk/ox_scores.php and the new scoring system can be found at http://phi.uhce.ox.ac.uk/pdf/OxfordScores/hip_score_guide.pdf. Free to use.

Method of administration. Self-administered; also used in postal surveys (24,31).

Scoring. According to the updated version, "Each of the 12 questions is scored in the same way with the score decreasing as the reported symptoms increase (i.e., become worse)" (22). Scores range from 0 to 4 (worst to best); see http://phi.uhce.ox.ac.uk/pdf/OxfordScores/hip_score_guide.pdf.

Computer scoring is not necessary. A maximum of 2 missing values can be accepted and replaced by mean value. Overall scores should not be calculated if more than 2 items are left unanswered. If 2 answers are indicated for 1 question, the worst response should be used for calculation of scores.

Score interpretation. According to the updated version, scores range from 0 to 48 (worst to best) (22). Cut off points based on large international data are under progress (22). Categories for the OHS based on data from the Harris Hip Score (HHS) and translated to the 0–48 scoring has suggested cut off scores: >41 as excellent, 34–41 as good, 27–33 as fair, and <27 as poor (32). Based on the original scoring system (12–60, best to worst) (21), <19 was excellent, 19–26 was good, 27–33 was fair, and >33 was poor (32).

According to the above classification by Kalairajah et al (32), the OHS at 6 months is a useful predictor of early revision after THR. A poor score was associated with a revision risk within 2 years of 7.6% for THR compared with risks of 0.7% for a good/excellent score (26). No normative values are available.

Respondent burden. The OHS takes between 2–15 minutes to complete (33).

Based on patient interviews, there were issues raised concerning item clarity and double-barreled questions (33,34).

Administrative burden. The OHS is a patient-reported questionnaire. Time to score is short, just sum items up. No training to score is necessary.

Translations/adaptations. Dutch (35), Japanese (27), German (36), and French (37) versions have been developed and evaluated. The OHS is widely used in many countries even though published validation studies are lacking. The Oxford Orthopaedic Scores also include a similar questionnaire for assessing outcome after knee replacement surgery (the Oxford Knee Score) together with questionnaires assessing shoulder surgery and shoulder instability.

Psychometric Information

Method of development. Questions were based on patient interviews where hip OA patients were asked to report their experience and problems. Patients were involved in face and content validity of the questionnaire (21). The OHS includes only 1 scale. OHS underwent item-response theory testing in 2004 by Fitzpatrick et al, and there was an overall good item fit of the data to the Rasch model (38).

Acceptability. Ninety percent of 6,174 questionnaires had no missing items. Most problems referred to item 6 (distance walked before severe pain). Older patients and patients with more severe medical problems were less likely to complete the questionnaire fully compared with younger and healthier patients (31). Ceiling effects (13.5%) were present in postoperatively collected data, but there were very low levels of floor effects (39,40).

Reliability. Internal consistency was measured in patients pre- and postsurgery; Cronbach's alpha varied between 0.84–0.93 (3, 6, 12, and 24 months) (21,31,35). Reproducibility was measured by the coefficient of repeatability according to the method of Bland and Altman, and found to be acceptable (21,35).

Validity. Developing the OHS, patients were asked to comment on and to include hip-related problems not addressed by the questionnaire for content validity (21). No hypotheses prior to analysis were provided measuring construct validity. Higher correlations to measures of pain and function than to psychological measures have been established (21,29,32,39,40). High correlation (r_s = 0.7, $P < 0.001$) was found between OHS and the HHS in THR patients (32).

Ability to detect change. OHS had greater responsiveness compared with generic measures (Short Form 36 and EuroQol 5-domain) and the disease-specific measures, the Western Ontario and McMaster Universities Osteoarthritis Index and the Arthritis Impact Measurement Scales. Effect size of the OHS varied between 2.38–3.1 at 6–24 months after THR (21,29,31,35,39–41) and was 1.84 at 6 months after revision surgery (41). According to Murray et al, the minimum clinically important difference can be expected to be between 3–5 points concerning joint replacement, but work is in progress (22).

A similar concept to the Patient Acceptable Symptom State was performed by Arden et al, relating the OHS to patient satisfaction with surgery after 12 and 24 months (42). The authors found that scores of 38 and 33 were associated with patient satisfaction at 12 and 24 months, respectively. However, the threshold varied according to preoperative scores and to body mass index limiting the clinical use of the threshold value.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The OHS assesses pain and function outcomes in patients undergoing hip replacement. It has shown acceptable to excellent psychometric properties and has been reported to be a useful predictor of early revision after THR.

Caveats and cautions. Like many of these questionnaires, the OHS has a few double-barreled questions that can be a problem to the patient. Questions have also been raised about the lack of items concerning activities requiring a large angle of hip flexion, as well as aids and medication; this information has to be addressed by other means.

Clinical usability. The questionnaire is easy to use due to self-administered distribution, and it only takes a few minutes to complete. A single administration will not provide much information on an individual, but repeated administrations might give some information

Research usability. The OHS was developed to supplement other generic outcome measures in systematic studies of hip replacement surgery with long-time followup; it is also feasible for surveys by post. Due to its shortness, the OHS questionnaire yields a high response rate and is therefore preferred for larger studies (24).

LEQUESNE INDEX OF SEVERITY FOR OSTEOARTHRITIS OF THE HIP (LISOH)

Description

Purpose. The LISOH was developed to evaluate the severity of hip osteoarthritis (OA) in drug trials in an adult French population, the long-term treatment effects for hip OA, and as a help in decision making regarding the need for hip replacement (43). The index covers OA-specific symptoms and physical functional disability (43–45). Developed in France in the early 1980s as an interview format to use in clinical drug trials, the instrument is available currently in several versions: interview based (43), self-administered (46), and in modified versions due to changed scoring and wording (45).

Content. A composite measure aggregating symptoms and function, which are not graded separately, where pain is analyzed by 5 items, maximum distance walked by 2 items, and activities of daily living (ADL) by 4 items.

Number of items. There are 11 items; the score ranges from 0 (no pain or no disability) to 24 (maximum pain or maximum disability) and is scored as the sum of all questions.

Recall period for items. Not specified.

Examples of use. To assess the severity of hip OA (47), the effectiveness of pharmacologic interventions (44), and to help with indications for surgery (46,48).

Practical Application

How to obtain. URL: http://www.oarsi.org/pdfs/pain_indexes/Lequesne_index.pdf. Free to use.

Method of administration. Patient-, interviewer-, or clinician-completed.

Scoring. The score ranges from 0–8 for each part (pain or discomfort, maximum distance walked, and ADL) resulting in a total score ranging from 0 to 24. The index was modified in 1991 when a question for sexual activity was included if appropriate, resulting in a maximum score of 28.

Score interpretation. Score 0–24 points (lower score indicates less dysfunction) where 0 = no handicap, 1–4 = mild handicap, 5–7 = moderate handicap, 8–10 = severe handicap, 11–13 = very severe handicap, and ≥14 = extremely severe handicap. A score over 11-12 points after appropriate treatment is suggested to indicate surgery (45). A sore >10 indicated a relative risk of 2.59 for total hip arthroplasty (48). The questions are suggested to score disabilities connected with a single hip. There are no indications of how to score in case of bilateral hip OA, complicating interpretation (47).

Respondent burden. Takes 2–5 minutes to complete (47,49,50).

Administrative burden. Some training may be needed for use of the interview-based questionnaire to reach interobserver reproducibility (43). Scoring takes only a few minutes (49).

Translations/adaptations. Validated for hip OA, it is available in French (original), English (47), German (46), Turkish (51), and Korean (52) but used in many languages where a published cultural adaptation is hard to find. Several cultural adaptations and validations have also been performed for the version used in knee OA.

Psychometric Information

Method of development. Developed in the early 1980s by specialists. Rasch analysis has been applied later in validity studies and has questioned the psychometric properties of the questionnaire (47).

Acceptability. Two of 10 patients needed some explanation to fill out the questionnaire in a French study using the Lequesne Index of Severity for Osteoarthritis of the Knee (49). In a postal survey, the constituent item response rate was 71% for the LISOH, which was lower than for the Short Form 36 (SF-36; 76–96%), but higher than the SF-36 total score (58%) (47).

Reliability. Satisfactory internal consistency for the composite score (alpha 0.83–0.84) has been presented (46,47,52). However, internal consistency was lower for the pain section compared with the function section (Cronbach's alpha 0.63 versus 0.84) (46). Recommendations are to only use the LISOH for group comparisons. Factor analysis did not show unidimensionality of the scale (47). Satisfactory test–retest reliability was found for the composite score, intraclass correlation coefficient 0.94 (46). For interrater reliability, the interview-based questionnaire had a mean deviation of 0.55 points when rated by 2 observers (43).

Validity. Doubtful construct validity (20,46,47). Also, the convergent validity of the questionnaire has been questioned (47).

Ability to detect change. Information concerning responsiveness is lacking. Active drug treatment has shown an effect size of 1.3–1.8 (45).

Critical Appraisal of Overall Value to the Rheumatology Community

Caveats and cautions. Problems raised are due to lack of validity, and the LISOH cannot be recommended for use as the single measure, neither in the clinic nor in research.

Clinical usability. Psychometric evaluations do not support the interpretation of scores on an individual level.

Research usability. Suggestions of more appropriate questionnaires for evaluation of pain and physical disability have been published in the last 10 years.

AMERICAN ACADEMY OF ORTHOPEDIC SURGEONS (AAOS) HIP AND KNEE QUESTIONNAIRE

Description

Purpose. The hip and knee core scale assesses hip and knee conditions and treatment improvements. The hip and knee questionnaire belongs to a series of lower extremity questionnaires initiated and developed by the AAOS. Intended for use in patients age ≥18 years. The original version was published in 2004 (53).

Content. The questionnaire covers stiffness, swelling, and pain in conjunction to functioning (walking on flat surfaces, going up or down stairs, lying in bed at night, ability to get around, and difficulties with taking on and off socks/stockings).

Number of items. 7 items, no subscales. If both hips are involved, the questions should be answered for the worse side.

Response options/scale. Likert scales with 5–7 response options (best to worse). Five response options for swelling and stiffness. Seven response options for pain and function, including 1 option for "could not do for other reasons," 7 options for getting around, and 6 response options for taking on and off socks/stockings.

Recall period for items. During the past week.

Examples of use. To measure functional impairment in patients treated for slipped capital femoral epiphysis (54).

Practical Application

How to obtain. Questionnaires and scoring instructions can be found at the AAOS web site: www.aaos.org/research/outcomes/outcomes_lower.asp.

Method of administration. Patient-administered questionnaire.

Scoring. Scoring includes both standardized and normative scores. Scoring instructions and a scoring worksheet can be obtained at the AAOS web site. Computer scoring is not necessary, but it speeds up the scoring process. Scores cannot be calculated if more than half of the items are missing.

Score interpretation. Standardized scores ranges from 0–100 (most disability to least disability). Standardized scores can then be transformed to normative scores using the mean and SD from the general healthy population. A patient scoring >50 on the normative scale will be above the general healthy population's average and a scoring <50 is under the general healthy population's average (55).

Respondent burden. The questionnaire takes only a few minutes to complete.

Administrative burden. Takes only a few seconds to score if the scoring sheet is used. If scored by hand, it takes ~15 minutes to score.

Translations/adaptations. There are versions for the lower extremity, for global sports/knee, and for the foot and ankle (53).

Psychometric Information

Method of development. In 1994, a consensus meeting was held and domains relevant for the lower extremity instruments were identified by group technique. The groups included clinicians and health-service researchers with an expertise in the field (53). The items in the scale were reduced from 28 to 7 due to factor analysis showing a considerable overlap with the Short Form 36 (SF-36) physical function scale (53).

Acceptability. Not studied.

Reliability. Internal consistency for patients with hip/knee diagnosis (n = 43) resulted in a Cronbach's alpha of 0.80. Test–retest was performed on 40 subjects and analyzed with the Pearson's correlation coefficient (r = 0.91) (53).

Validity. Face and content validity were determined by the item selection process. Construct validity of the hip/knee scale was performed by analyzing data from 43 patients in the hip/knee group, yielding correlations of 0.95 with the lower extremity core scale, 0.70 with the unweighted mean of the SF-36 physical health score, 0.73 and 0.69 with physician assessment of function and pain, respectively. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) was assessed for criterion validity, a global score for the WOMAC was calculated, and the correlation with the hip/knee core score was 0.89 (53).

Ability to detect change. Differences between change scores were not calculated for the hip/knee core scale, but they were for the lower extremity core scale after 24 months. Change scores on the lower extremity questionnaire were correlated with a patient-physician–generated score regarding the perception of improvement during the last year (r = 0.53).

In a regression analysis with the transition score generated from patient-physician perception of improvement as dependant variable, the lower extremity core scale accounted for 40% of the variance, which was the highest among the tested outcome measures (53).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The AAOS lower extremity questionnaires went through a psychometric evaluation reported by Johanson et al in 2004 (53); this is, however, the only one performed and published. The authors conclude that the measures combined with the SF-36 will provide useful information concerning orthopedic outcome in patients with lower extremity diagnoses. The usefulness of the questionnaire will need to be studied further.

Clinical usability. Developed for use in the clinic as well as in research. The transformation of standardized scores to normative scores can be useful also in the clinic. Further testing of the instrument is warranted.

AUTHOR CONTRIBUTIONS

Both authors were involved in drafting the article or revising it critically for important intellectual content, and both authors approved the final version to be published.

REFERENCES

1. Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. An end-result study using a new method of result evaluation. J Bone Joint Surg Am 1969; 51:737–55.
2. Garellick G, Malchau H, Herberts P. Specific or general health outcome measures in the evaluation of total hip replacement: a comparison between the Harris Hip Score and the Nottingham Health Profile. J Bone Joint Surg Br 1998;80:600–6.
3. Lieberman JR, Dorey F, Shekelle P, Schumacher L, Kilgus DJ, Thomas BJ, et al. Outcome after total hip arthroplasty: comparison of a traditional disease-specific and a quality-of-life measurement of outcome. J Arthroplasty 1997;12:639–45.
4. Soderman P, Malchau H, Herberts P. Outcome of total hip replacement: a comparison of different measurement methods. Clin Orthop Relat Res 2001;390:163–72.
5. Frihagen F, Grotle M, Madsen JE, Wyller TB, Mowinckel P, Nordsletten L. Outcome after femoral neck fractures: a comparison of Harris Hip Score, Eq-5d and Barthel Index. Injury 2008;39:1147–56.
6. Hoeksma HL, van den Ende CH, Ronday HK, Heering A, Breedveld FC. Comparison of the responsiveness of the Harris Hip Score with generic measures for hip function in osteoarthritis of the hip. Ann Rheum Dis 2003;62:935–8.
7. Wamper KE, Sierevelt IN, Poolman RW, Bhandari M, Haverkamp D. The Harris Hip Score: do ceiling effects limit its usefulness in orthopedics? Acta Orthop 2010;81:703–7.
8. Soderman P, Malchau H. Is the Harris Hip Score system useful to study the outcome of total hip replacement? Clin Orthop Relat Res 2001;384: 189–97.
9. Kirmit L, Karatosun V, Unver B, Bakirhan S, Sen A, Gocen Z. The reliability of hip scoring systems for total hip arthroplasty candidates: assessment by physical therapists. Clin Rehabil 2005;19:659–61.
10. Shi HY, Mau LW, Chang JK, Wang JW, Chiu HC. Responsiveness of the Harris Hip Score and the SF-36: five years after total hip arthroplasty. Qual Life Res 2009;18:1053–60.
11. Soderman P. On the validity of the results from the Swedish National Total Hip Arthroplasty Register. Acta Orthop Scand Suppl 2000;71:1–33.
12. Klassbo M, Larsson E, Mannevik E. Hip Disability and Osteoarthritis Outcome Score: an extension of the Western Ontario and McMaster Universities Osteoarthritis Index. Scand J Rheumatol 2003;32:46–51.
13. Nilsdotter AK, Lohmander LS, Klassbo M, Roos EM. Hip Disability and Osteoarthritis Outcome Score (HOOS): validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003;4:10.
14. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833–40.
15. Davis AM, Perruccio AV, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16: 551–9.
16. Gossec L, Paternotte S, Maillefert JF, Combescure C, Conaghan PG, Davis AM, et al. The role of pain and functional impairment in the decision to recommend total joint replacement in hip and knee osteoarthritis: an international cross-sectional study of 1909 patients: report of the OARSI-OMERACT Task Force on Total Joint Replacement. Osteoarthritis Cartilage 2011;19:147–54.
17. De Groot IB, Reijman M, Terwee CB, Bierma-Zeinstra SM, Favejee M, Roos EM, et al. Validation of the Dutch version of the Hip Disability and Osteoarthritis Outcome Score. Osteoarthritis Cartilage 2007;15:104–9.
18. Ornetti P, Parratte S, Gossec L, Tavernier C, Argenson JN, Roos EM, et al. Cross-cultural adaptation and validation of the French version of the Hip Disability and Osteoarthritis Outcome Score (HOOS) in hip osteoarthritis patients. Osteoarthritis Cartilage 2010;18:522–9.
19. Thorborg K, Roos EM, Bartels EM, Petersen J, Holmich P. Validity, reliability and responsiveness of patient-reported outcome questionnaires when assessing hip and groin disability: a systematic review. Br J Sports Med 2010;44:1186–96.
20. Veenhof C, Bijlsma JW, van den Ende CH, van Dijk GM, Pisters MF, Dekker J. Psychometric evaluation of osteoarthritis questionnaires: a systematic review of the literature. Arthritis Rheum 2006;55:480–92.
21. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. J Bone Joint Surg Br 1996;78:185–90.
22. Murray DW, Fitzpatrick R, Rogers K, Pandit H, Beard DJ, Carr AJ, et al. The use of the Oxford Hip and Knee Scores. J Bone Joint Surg Br 2007;89:1010–4.
23. Baker PN, van der Meulen JH, Lewsey J, Gregg PJ. The role of pain and function in determining patient satisfaction after total knee replacement: data from the National Joint Registry for England and Wales. J Bone Joint Surg Br 2007;89:893–900.
24. Dunbar MJ, Robertsson O, Ryd L, Lidgren L. Appropriate questionnaires for knee arthroplasty: results of a survey of 3600 patients from The Swedish Knee Arthroplasty Registry. J Bone Joint Surg Br 2001; 83:339–44.
25. Pynsent PB, Adams DJ, Disney SP. The Oxford Hip and Knee outcome questionnaires for arthroplasty. J Bone Joint Surg Br 2005;87:241–8.
26. Rothwell AG, Hooper GJ, Hobbs A, Frampton CM. An analysis of the Oxford Hip and Knee Scores and their relationship to early joint revision in the New Zealand Joint Registry. J Bone Joint Surg Br 2010;92: 413–8.
27. Uesugi Y, Makimoto K, Fujita K, Nishii T, Sakai T, Sugano N. Validity and responsiveness of the Oxford Hip Score in a prospective study with Japanese total hip arthroplasty patients. J Orthop Sci 2009;14:35–9.
28. Wylde V, Blom AW, Whitehouse SL, Taylor AH, Pattison GT, Bannister GC. Patient-reported outcomes after total hip and knee arthroplasty: comparison of midterm results. J Arthroplasty 2009;24:210–6.
29. Dawson J, Fitzpatrick R, Frost S, Gundle R, McLardy-Smith P, Murray D. Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. J Bone Joint Surg Br 2001;83:1125–9.
30. Field RE, Cronin MD, Singh PJ. The Oxford Hip Scores for primary and revision hip replacement. J Bone Joint Surg Br 2005;87:618–22.
31. Fitzpatrick R, Morris R, Hajat S, Reeves B, Murray DW, Hannen D, et al. The value of short and simple measures to assess outcomes for patients of total hip replacement surgery. Qual Health Care 2000;9:146–50.
32. Kalairajah Y, Azurza K, Hulme C, Molloy S, Drabu KJ. Health outcome measures in the evaluation of total hip arthroplasties: a comparison between the Harris Hip Score and the Oxford Hip Score. J Arthroplasty 2005;20:1037–41.
33. McMurray R, Heaton J, Sloper P, Nettleton S. Measurement of patient perceptions of pain and disability in relation to total hip replacement: the place of the Oxford Hip Score in mixed methods. Qual Health Care 1999;8:228–33.
34. Wylde V, Learmonth ID, Cavendish VJ. The Oxford Hip Score: the patient's perspective. Health Qual Life Outcomes 2005;3:66.
35. Gosens T, Hoefnagels NH, de Vet RC, Dhert WJ, van Langelaan EJ, Bulstra SK, et al. The "Oxford Heup Score": the translation and validation of a questionnaire into Dutch to evaluate the results of total hip arthroplasty. Acta Orthop 2005;76:204–11.
36. Naal FD, Sieverding M, Impellizzeri FM, von Knoch F, Mannion AF, Leunig M. Reliability and validity of the cross-culturally adapted German Oxford Hip Score. Clin Orthop Relat Res 2009;467:952–7.
37. Delaunay C, Epinette JA, Dawson J, Murray D, Jolles BM. Cross-cultural

adaptations of the Oxford-12 Hip score to the French speaking population. Orthop Traumatol Surg Res 2009;95:89–99.

38. Fitzpatrick R, Norquist JM, Jenkinson C, Reeves BC, Morris RW, Murray DW, et al. A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. Qual Life Res 2004;13:331–8.
39. Garbuz DS, Xu M, Sayre EC. Patients' outcome after total hip arthroplasty: a comparison between the Western Ontario and McMaster Universities Index and the Oxford 12-item Hip Score. J Arthroplasty 2006;21:998–1004.
40. Ostendorf M, van Stel HF, Buskens E, Schrijvers AJ, Marting LN, Verbout AJ, et al. Patient-reported outcome in total hip replacement: a comparison of five instruments of health status. J Bone Joint Surg Br 2004;86:801–8.
41. Dawson J, Fitzpatrick R, Murray D, Carr A. Comparison of measures to assess outcomes in total hip replacement surgery. Qual Health Care 1996;5:81–8.
42. Arden NK, Kiran A, Judge A, Biant LC, Javaid MK, Murray DW, et al. What is a good patient reported outcome after total hip replacement? Osteoarthritis Cartilage 2011;19:155–62.
43. Lequesne MG, Mery C, Samson M, Gerard P. Indexes of severity for osteoarthritis of the hip and knee: validation—value in comparison with other assessment tests. Scand J Rheumatol Suppl 1987;65:85–9.
44. Lequesne M. Indices of severity and disease activity for osteoarthritis. Semin Arthritis Rheum 1991;20 Suppl 2:48–54.
45. Lequesne MG. The algofunctional indices for hip and knee osteoarthritis. J Rheumatol 1997;24:779–81.
46. Stucki G, Sangha O, Stucki S, Michel BA, Tyndall A, Dick W, et al. Comparison of the WOMAC (Western Ontario and McMaster Universities) Osteoarthritis Index and a self-report format of the self-administered Lequesne-Algofunctional Index in patients with knee and hip osteoarthritis. Osteoarthritis Cartilage 1998;6:79–86.
47. Dawson J, Linsell L, Doll H, Zondervan K, Rose P, Carr A, et al. Assessment of the Lequesne Index of Severity for Osteoarthritis of the Hip in an elderly population. Osteoarthritis Cartilage 2005;13:854–60.
48. Dougados M, Gueguen A, Nguyen M, Berdah L, Lequesne M, Mazieres B, et al. Requirement for total hip arthroplasty: an outcome measure of hip osteoarthritis? J Rheumatol 1999;26:855–61.
49. Faucher M, Poiraudeau S, Lefevre-Colau MM, Rannou F, Fermanian J, Revel M. Algo-functional assessment of knee osteoarthritis: comparison of the test–retest reliability and construct validity of the WOMAC and Lequesne indexes. Osteoarthritis Cartilage 2002;10:602–10.
50. Lequesne MG, Samson M. Indices of severity in osteoarthritis for weight bearing joints. J Rheumatol Suppl 1991;27:16–8.
51. Basaran S, Guzel R, Seydaoglu G, Guler-Uysal F. Validity, reliability, and comparison of the WOMAC Osteoarthritis Index and Lequesne Algofunctional Index in Turkish patients with hip or knee osteoarthritis. Clin Rheumatol 2010;29:749–56.
52. Bae SC, Lee HS, Yun HR, Kim TH, Yoo DH, Kim SY. Cross-cultural adaptation and validation of Korean Western Ontario and McMaster Universities (WOMAC) and Lequesne osteoarthritis indices for clinical research. Osteoarthritis Cartilage 2001;9:746–50.
53. Johanson NA, Liang MH, Daltroy L, Rudicel S, Richmond J. American Academy of Orthopaedic Surgeons lower limb outcomes assessment instruments: reliability, validity, and sensitivity to change. J Bone Joint Surg Am 2004;86-A:902–9.
54. DeLullo JA, Thomas E, Cooney TE, McConnell SJ, Sanders JO. Femoral remodeling may influence patient outcomes in slipped capital femoral epiphysis. Clin Orthop Relat Res 2007;457:163–70.
55. Hunsaker FG, Cioffi DA, Amadio PC, Wright JG, Caughlin B. The American Academy of Orthopedic Surgeons outcomes instruments: normative values from the general population. J Bone Joint Surg Am 2002;84-A:208–15.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S287–S296
DOI 10.1002/acr.20544

Measures of Sleep in Rheumatologic Diseases

Epworth Sleepiness Scale (ESS), Functional Outcome of Sleep Questionnaire (FOSQ), Insomnia Severity Index (ISI), and Pittsburgh Sleep Quality Index (PSQI)

THEODORE A. OMACHI

INTRODUCTION

Fatigue is a major symptom associated with rheumatologic diseases such as systemic lupus erythematosus and rheumatoid arthritis and may be a direct manifestation of disease activity; however, such fatigue may also be related to sleep disturbances (1,2). Indeed, sleep disturbances are common in a variety of rheumatologic diseases (3–5). Such disturbed sleep may be due to pain, depression, lack of exercise, or corticosteroid usage (6–8). Sleep quality may also be impaired by comorbid sleep disorders, such as obstructive sleep apnea or restless legs syndrome, the prevalences of which are reported to be high in rheumatologic populations (9–12). Sleep disturbances may, in turn, impact functional disability, lower pain thresholds, or impair immune function and therefore contribute to rheumatologic-associated morbidities (13–15). Sleep disturbances in fibromyalgia and rheumatoid arthritis have received relatively more attention than in other rheumatologic disease; however, even in fibromyalgia and rheumatoid arthritis, there are many unanswered questions related to the causes and outcomes of sleep disturbances (3).

The study of sleep disturbances can be onerous because gold standard direct tests, such as polysomnography and multiple sleep latency testing, are both expensive and require considerable commitment of time from research subjects. Laboratory-based sleep studies may present an additional challenge in rheumatologic populations in whom mobility restriction and pain may significantly increase subject burden. Therefore, there is strong impetus for utilizing patient-reported measures in assessing sleep and sleep-related outcomes in rheumatologic diseases.

Four patient-reported measures are discussed in this section, each of which captures a different sleep-related domain and has been extensively utilized in a variety of populations: 1) the Epworth Sleepiness Scale, which assesses daytime sleepiness, 2) the Functional Outcome of Sleep Questionnaire, which assesses sleep-related quality of life, 3) the Insomnia Severity Index, which measures the subjective symptoms and consequences of difficulties initiating and maintaining sleep, and 4) the Pittsburgh Sleep Quality Index, which more generally assesses perceived sleep quality. Please note that the Medical Outcomes Study Sleep Scale, a global measure of sleep quality and sleep-related outcomes, is discussed separately in the Fibromyalgia Section of this issue. None of the scales reviewed here were developed specifically for rheumatologic or musculoskeletal conditions and, indeed, each has relied heavily on populations with primary sleep disorders for validation. To varying extents, as discussed below, each of these measures has been used in rheumatologic populations. Nonetheless, clinicians and researchers must carefully consider their objectives and the appropriateness of their populations in selecting a sleep questionnaire to meet their needs.

EPWORTH SLEEPINESS SCALE (ESS)

Description

Purpose. To measure daytime sleepiness (16).

Content. The ESS is intended to measure the single factor of "somnoficity." The instrument asks subjects to rate "in recent times" how likely they would be to "doze off or fall asleep" in 8 different common situations of daily living, such as "sitting and reading" or "watching TV." The ESS asks respondents to "try to work out how they would have affected you," even if they have not done a given activity recently.

Number of items. 8 items.

Response options/scale. The questionnaire has a 4-point Likert response format (0 = would never doze, 1 = slight chance of dozing, 2 = moderate chance of dozing, and 3 = high chance of dozing).

Recall period for items. "Recent times." Further specificity is not provided.

Endorsements. No.

Dr. Omachi's work was supported by the NIH/National Heart, Lung, and Blood Institute (grant K23-HL-102159).

Theodore A. Omachi, MD, MBA: University of California, San Francisco.

Address correspondence to Theodore A. Omachi, MD, MBA, Division of Pulmonary, Critical Care, and Sleep Medicine, University of California, San Francisco, Sleep Disorders Center, 2330 Post Street, Suite 420, San Francisco, CA 94115. E-mail: omachi@ucsf.edu.

Submitted for publication January 31, 2011; accepted in revised form June 23, 2011.

Examples of use. The ESS has been used frequently in studies of obstructive sleep apnea (OSA), but has also been applied to study sleepiness related to Parkinson's disease (17), multiple sclerosis (18), asthma (19), gastroesphogeal reflux (20), and multiple other chronic diseases. Its usage in the rheumatologic literature has been more limited than in primary sleep disorders, but it has been applied in examining the effects of chronic pain on sleepiness (21, 22).

Practical Application

How to obtain. The survey instrument is available in the original validating publication (16), and is also available at http://epworthsleepinessscale.com. An annual license fee may be applicable if usage is "deemed commercial in nature." Permission to use can be obtained from Murray W. Johns, PhD, who can be contacted through the above web site or at Epworth Sleep Centre, Melbourne, Victoria, Australia. E-mail: mjohns@optalert.com.

Method of administration. Written survey instrument.

Scoring. The 8 Likert response items are summed to calculate a total score.

Score interpretation. Score range is 0–24, with higher scores indicating greater daytime sleepiness. Scores ≥11 are generally considered abnormal, or positive for excessive daytime sleepiness (EDS). This criteria for EDS was based on a mean ± SD score of 4.5 ± 2.8 among 72 healthy Australian workers (23).

Respondent burden. 2–3 minutes.

Administration burden. Time to score is <1 minute.

Translations/adaptations. The ESS has been translated and validated in multiple languages, including Spanish, German, Mandarin Chinese, Turkish, and Greek (24–28).

Psychometric Information

Method of development. The 8 situations assessed for likelihood of falling asleep were selected based on earlier research regarding low-stimulating environments that were likely to be soporific (29).

Acceptability. Item-response rates are reported to be high, with Johns and Hocking reporting <1% of surveys having missing data (23). In a recent study, score distributions were reasonably normal among community-dwelling US adults, with a mean ± SD score of 8.2 ± 3.9 (30).

Reliability. There was adequate internal consistency with Cronbach's alpha (range $\alpha = 0.74-0.88$) (31,32). Test–retest reliability was reported to be high based on testing separated in time by 5 months in healthy subjects ($r = 0.82$, $P < 0.001$) (31). In subjects with OSA, with testing separated by an average of 71 days, $r = 0.73$ ($P < 0.001$) (33).

Validity. Concurrent validity of the ESS has been assessed as its correlation with mean sleep latency on multiple sleep latency tests (MSLT) in which subjects are asked to take a series of brief naps over the course of several hours. In such studies, the ESS showed correlations in the expected directions of between 0.30 and 0.37 (34, 35). Although this correlation is not exceptionally high, the validity of the ESS has also been argued based on evidence that it predicts, better than MSLT, the presence of narcolepsy, a condition which is by definition associated with excessive daytime somnolence (36). The validity of the ESS has also been established based on its association with the Respiratory Disturbance Index among OSA patients, and its responsiveness to treatment in OSA (16,31).

Ability to detect change. Based on results from clinical trials, the ESS is sensitive to change, with therapies thought to reduce sleepiness, showing improvements in ESS (17,18,37). Minimally clinical important differences are not reported.

Discussion

The ESS is one of the most widely used measures, both clinically and in sleep medicine research, with the original validation article having been referenced more than 3,000 times in peer-reviewed publications. Its attractiveness is based in part on its ease of administration, as well as the simplicity of the concept it is measuring, daytime sleepiness. Although the MSLT is considered by many to be the gold-standard for measuring sleepiness (34), it is often not practical for research or clinical purposes. By specifically asking about the likelihood of falling asleep in various situations, rather than the effects of sleepiness on daily activities, the ESS may hold some theoretical advantages in distinguishing fatigue from sleepiness, where fatigue is defined as a subjective lack of physical or mental energy to carry out desired activities (38). This may be important in rheumatologic diseases, which might be expected to cause significant fatigue independent of sleepiness, although the application of the ESS to rheumatologic conditions has been relatively limited, and validation of this distinction has not been established. An additional caution is that the ESS cannot distinguish between sleepiness as a result of disturbed sleep and sleepiness resulting from other causes, such as medication effects.

FUNCTIONAL OUTCOMES OF SLEEP QUESTIONNAIRE (FOSQ)

Description

Purpose. To assess the impact of excessive sleepiness on functional outcomes relevant to daily behaviors and sleep-related quality of life (39).

Content. The instrument asks subjects if they have had difficulty performing specific activities because of "being sleepy or tired." It provides instructions to respondents informing them that the words "sleepy" and "tired" mean "the feeling that you can't keep your eyes open, your head is droopy, that you want to 'nod off,' or that you feel the urge to take a nap. These words do not refer to the tired or fatigued feeling you may have after you have exercised."

In 30 items, the FOSQ then assesses difficulty, due to sleepiness, in performing activities of daily living and recreational activities, which are categorized into the following 5 subscales: 1) activity level (9 items), 2) vigilance (7 items), 3) intimacy and sexual relationships (4 items), 4) general productivity (8 items), and 5) social outcomes

(2 items). A shorter 10-item version, the FOSQ-10, was published in 2009 using selected items from each subscale and providing the same definition of sleepy and tired (40). Items for the FOSQ-10 are distributed among the same subscales as follows: 1) activity level (3 items), 2) vigilance (3 items), 3) intimacy and sexual relationships (1 item), 4) general productivity (2 items), and 5) social outcomes (1 item). However, because of the limited number of items in each subscale for the FOSQ-10, the authors recommend that only the total score for the FOSQ-10 be utilized, rather than individual subscales.

Number of items. There are 30 items in the original FOSQ-30, and 10 items in the FOSQ-10.

Response options/scale. The questionnaire has a 4-point Likert response format (e.g., 1 = extreme difficulty, 2 = moderate difficulty, 3 = a little difficulty, and 4 = no difficulty). A response alternative is also available for respondents to indicate that they do not engage in the activity for reasons other than being sleepy or tired.

Recall period for items. Not specified. Question stems imply current difficulty.

Endorsements. No.

Examples of use. The FOSQ-30 has been used to assess response to therapies in randomized clinical trials (37,41,42) or prospective cohort studies (43) and to assess the impact of known or suspected sleep disturbances on daytime function (44–48). For example, Burke et al report that although opioids-dependent individuals reported significant sleep disturbance, such sleep disturbance did not appear to affect daily functioning as assessed by the FOSQ (45). The FOSQ has been applied to a limited extent in populations with rheumatologic disease (49,50). The FOSQ is frequently used as a measure of sleep-specific health-related quality of life.

Practical Application

How to obtain. Available from the authors. Permission for use is required. Contact Terri E. Weaver, PhD, RN, University of Illinois at Chicago, 845 South Damen Avenue, MC 802, Chicago, IL 60612. E-mail: teweaver@uic.edu.

Method of administration. Self-administered written questionnaire.

Scoring. For both the FOSQ-30 and FOSQ-10, an average score is calculated for each subscale, and the 5 subscales are totaled to produce a total score. Missing responses, and responses from activities in which the respondent does not participate regularly "for reasons other than being sleepy or tired," are not included in the score calculation (i.e., not included in the calculation of average value for subscales). Therefore, missing responses do not necessarily prevent score calculation. Subscale scores for both the FOSQ-10 and FOSQ-30 range from 1–4 with total scores ranging from 5–20.

Score interpretation. Score range is 5–20 points, with higher scores indicating better functional status.

Respondent burden. The FOSQ is written at a fifth-grade reading level. Time to complete the FOSQ-30 is reported to be 15 minutes (39). Time to complete the FOSQ-10 is not reported. Although the FOSQ-10 has one-third the number of questions, it may take longer than one-third of the time of the FOSQ-30 to administer, given that the length of instructions related to defining sleepy and tired are unchanged.

Administration burden. Time to score is not reported, but is estimated here to be ~3–5 minutes if done by hand.

Translations/adaptations. The FOSQ-30 has been translated and validated in peer-reviewed publications in multiple languages including Spanish, German, Turkish, and Norwegian (51–55). Multiple other translated versions of the FOSQ-30, although not specifically validated in peer-reviewed publications, are also available from the authors.

Psychometric Information

Method of development. Based on Granger's model of disability, 74 items were originally identified and tested in 3 distinct cohorts, consisting largely of participants with either confirmed sleep apnea or those referred to sleep disorders clinics. Forty-four items were then eliminated because 1) a high level of agreement between questions about degree of difficulty and frequency of symptoms lead to elimination of questions about frequency of symptoms, 2) certain items reduced the reliability (Cronbach's alpha) of the subscales and were therefore eliminated, and 3) items which did not meet the loading criterion of >0.40 were eliminated.

Acceptability. Information on the number of missing items was not reported in original FOSQ development, although a given respondent's total score and subscale scores are not invalidated by missing items. Scores may cluster toward the high-end of the FOSQ range (scores 5–20), especially in populations selected from the community or without sleep complaints. Among older community-dwelling adults, Gooneratne et al report that the mean ± SD FOSQ total score was 19.29 ± 0.67 among subjects without excessive daytime sleepiness (EDS; based on Epworth Sleepiness Scale scores) and was 17.91 ± 2.00 among subjects with EDS (56). Nonresponse may be a problem for questions related to intimacy and sexual activity, since a majority of respondents in that study did not answer these questions (56).

Reliability. In their original development paper, Weaver et al report a high internal consistency with Cronbach's alpha ($\alpha = 0.95$) for the 30-item FOSQ, after elimination of items that reduced the Cronbach's alpha (39). For the FOSQ-10, Cronbach's alpha was $\alpha = 0.87$ (40). Test–retest reliability for the FOSQ-30 was high, based on testing separated by 1 week without interval intervention ($r = 0.90$).

Validity. Concurrent validity of the FOSQ-30 was established based on moderate correlation with the Sickness Impact Profile (SIP), a general (not disease-specific) measure of functional status outcomes, and the Short Form 36 (SF-36) health survey. FOSQ subscales generally correlated more highly with related SIP and SF-36 subscales and less with unrelated SIP and SF-36 subscales. Discriminant validity was established based on differences in scores between respondents seeking evaluation for sleep disorders and individuals without sleep complaints (*t*-test −5.88, $P < 0.001$) (39).

The FOSQ-10 total score was robustly associated with the FOSQ-30 total score, (r = 0.96, $P < 0.0001$), explaining 92% of the variance of the longer version. The subscales of the FOSQ-10 and FOSQ-30 were also highly correlated with Pearson's correlation coefficient as r = 0.83–0.97 ($P < 0.0001$ for all) (40). Scores on the FOSQ-10 were also significantly lower in untreated sleep apnea patients (mean ± SD 12.48 ± 3.23) as compared to controls without sleep disorders (mean ± SD 17.81 ± 3.10) ($P < 0.0001$), suggesting discriminant validity.

Ability to detect change. Sensitivity to change has been demonstrated in clinical trials showing improvements in the FOSQ-30 resulting from therapies such as modafinil or positive airway pressure therapy (37,42). The FOSQ-10 has also shown improvements resulting from positive airway pressure therapy in patients with sleep apnea (40). Minimally clinical important differences are not reported.

Discussion

The FOSQ is a widely used measure of functional status resulting from sleepiness and has been effectively employed as a measure of sleep-related quality of life. It has been applied most often in the context of primary sleep disorders, sleep apnea in particular, but it is not specific for any particular disease. As with the Epworth Sleepiness Scale, the FOSQ cannot distinguish between impairment resulting from disturbed sleep or that due to medications such as opiates. The FOSQ has not specifically been validated in rheumatologic populations or applied widely in cohorts with rheumatologic disease. Nonetheless, investigators intending to determine the extent to which rheumatologic diseases impair HRQOL due to sleepiness or disturbed sleep may find the FOSQ to be a useful outcome, since many other measures of sleep-related HRQOL are specific to sleep apnea or primary sleep disorders (57). One strength of the FOSQ is its inquiry about items related to intimacy and sexual function, a subject area not captured in many instruments. However, nonresponse to these items may present a problem, as indicated in one study (56).

The FOSQ-10, a shorter version of the FOSQ, was published in 2009, and its total score and individual subscales correlated nicely with the FOSQ-30. Further validation and examples of implementation are not yet available, but this may be an appealing version if the FOSQ-30 is not practical because of length.

INSOMNIA SEVERITY INDEX (ISI)

Description

Purpose. To be a brief self-report instrument measuring self-perception of insomnia symptoms as well as the degree of concerns or distress caused by those symptoms.

Content. Content of the ISI corresponds in part to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnostic criteria for insomnia. In a 7-item questionnaire, with 1 item for each of the following categories, the ISI assesses 1) difficulty with sleep onset, 2) difficulty with sleep maintenance, 3) problem with early awakening, 4) satisfaction with sleep pattern, 5) interference with daily functioning as a result of sleep problems, 6) noticeability of sleep problem to others, and 7) degree of distress caused by sleep problem.

Number of items. 7 items.

Response options/scale. Each item has a 5-point Likert response format.

Recall period for items. Last 2 weeks.

Endorsements. No.

Examples of use. The ISI was developed to be an outcomes measure for insomnia research and has frequently been used as an outcome in clinical trials, both of pharmacologic therapies and behavioral interventions (58–64). It has also been used to identify morbidity and poor outcomes associated with insomnia, including in rheumatologic diseases (65,66).

Practical Application

How to obtain. The written questionnaire was published in the original validation study (67). Permission for usage can be obtained from the author. Contact Charles M. Morin, PhD, Université Laval and Centre de recherche Université Laval-Robert Giffard, Québec, Canada. E-mail: cmorin@psy.ulaval.ca.

Method of administration. Authors report that ISI is available in 3 forms: written questionnaire for self-administration, written questionnaire for significant other administration, and clinician administration. The self-administered version was the primary focus of validation (67), and this review also focuses on that version, except where otherwise noted.

Scoring. The 7 Likert response items are summed to determine total score.

Score interpretation. The score range is 0–28 points, with higher scores indicating greater insomnia severity. The suggested guidelines for score interpretation is 0–7 for no clinically significant insomnia, 8–14 for subthreshold insomnia, 15–21 for clinical insomnia (moderate severity), and 22–28 for clinical insomnia (severe). However, empiric validation of these guidelines is required. Savard et al recommend a cut off score of 8 for detection of sleep difficulties, which yielded a sensitivity of 94.7% and a specificity of 47.4% among cancer patients based on a gold standard of the Insomnia Interview Schedule, a semistructured interview based on DSM-IV criteria (68). Recommended cut off scores for other populations have not been well established empirically.

Respondent burden. Time to complete is <5 minutes.

Administration burden. Time to score is <1 minute.

Translations/adaptations. French-Canadian, Spanish, and Chinese versions have been validated (68–70). Only the clinician-administered version was validated in Chinese.

Psychometric Information

Method of development. Items for the ISI were selected based on DSM-IV and International Classification of Sleep Disorders criteria for insomnia. The ISI was based closely on the Sleep Impairment Index, an earlier measure developed by Morin (71,72).

Acceptability. A floor effect may be present in populations with low prevalence of insomnia symptoms. Among French-Canadian cancer patients, the mean ± SD ISI score was 7.3 ± 6.3 (68). However, among patients referred to a sleep clinic for insomnia, scores were less skewed with a mean ± SD score of 15.4 ± 4.2 (67). Among primary-care Chinese-speaking older adults, the mean ± SD score was 10.4 ± 5.2 (70). Information about missing items and educational attainment of subjects was not presented in validation studies (67).

Reliability. Adequate internal consistency is suggested by a Cronbach's alpha of $\alpha = 0.76$ at baseline in the original validation study, $\alpha = 0.81$ among community-dwelling older Chinese patients, and $\alpha = 0.90$ among French-Canadian cancer patients (67,68,70). Savard et al report that among cancer patients, the test–retest reliability (Pearson's correlation coefficient) is $r = 0.83$ ($P < 0.0001$) after 1 month, $r = 0.77$ ($P < 0.0001$) after 2 months, and $r = 0.73$ ($P < 0.0001$) after 3 months (68).

Validity. *Construct validity.* Because the ISI is based on DSM-IV criteria, it has good face validity. A principal component analysis yielded 3 components consistent with diagnostic criteria for insomnia (impact, severity, and satisfaction) that explained 72% of the total variance (67). Among cancer patients, 2 factors corresponding to severity and impact were identified (68).

Concurrent validity. Bastien and colleagues provided evidence for concurrent validity as correlation between ISI and sleep diary variables, where $r = -0.35$ ($P < 0.05$) at baseline for correlation between ISI and sleep efficiency (defined as percentage of time asleep when in bed), as recorded in a sleep diary over a period of 1–2 weeks. Correlation with sleep diary was higher after insomnia treatment, with $r = -0.60$ ($P < 0.05$). The ISI was not correlated with sleep efficiency as recorded on polysomnography (PSG) in a sleep laboratory over 3 consecutive nights ($r = 0.09$, P 0.05), although the ISI sleep onset item was correlated with time to sleep onset as recorded by PSG ($r = 0.45$, $P < 0.05$) (67).

Ability to detect change. *Sensitivity to change.* When comparing the change (pretreatment for insomnia versus posttreatment) in the ISI score, the correlation for ISI change was $r = -0.37$ ($P < 0.05$) as compared with the change in sleep efficiency recorded by sleep diary, and $r = -0.36$ ($P < 0.01$) as compared with change in sleep efficiency recorded in sleep laboratory on PSG (67). In trials of pharmacologic therapies for insomnia, the ISI has also demonstrated sensitivity to change. For example, in a 6-month randomized double-blind trial, the ISI declined among eszopiclone users, from mean ± SD 17.9 ± 4.1 at baseline to 8.3 ± 6.0 at 6 months. In the placebo group, the change in ISI score was mean ± SD 17.8 ± 4.1 at baseline and 12.9 ± 5.7 at 6 months ($P < 0.0001$ for difference between groups at 6 months).

Minimum clinically important difference (MCID). An MCID of 6 points has been recommended based on an analysis that demonstrated such an improvement in scores was associated with the following quality anchors: 48% reduction in likelihood of "feeling worn out" at 6 months (from the Short Form 36 Health Survey), 46% less likely to be "able to think clearly" (from the Work Limitations Questionnaire), and 52% less likely to report "feeling fatigued" (from the Fatigue Severity Scale). A 6-point change was equivalent to 1.5 SDs in this study (73).

Discussion

The ISI has high face validity, is a relatively short instrument, and has been used extensively in clinical research. It has been validated in a number of different cohorts, both those referred for insomnia symptoms, as well as cohorts selected outside of sleep referral centers. The suggested guidelines for classifying insomnia require further validation, and based on the research of Savard and colleagues, there does not appear to be a clear threshold above which clinical insomnia can be diagnosed with high certainty but below which it can also be excluded with confidence (68). Moreover, and particularly relevant to research in rheumatologic diseases, the instrument does not distinguish between causes of insomnia, whether psychophysiologic in origin or related to pain or other symptoms from medical comorbidity. Nonetheless, it has been used effectively in populations with comorbid disease, including cohorts with rheumatologic diseases, and is a useful and brief instrument.

PITTSBURGH SLEEP QUALITY INDEX (PSQI)

Description

Purpose. To measure sleep quality and disturbances over the prior month and to discriminate between "good" and "poor" sleepers (74).

Content. The PSQI consists of 7 components: subjective sleep quality (1 item), sleep latency (2 items), sleep duration (1 item), habitual sleep efficiency (3 items), sleep disturbances (9 items), use of sleeping medications (1 item), and daytime dysfunction (2 items).

Number of items. Nineteen items are included in scoring. Five additional items, to be completed by a bed partner, are included in the questionnaire and may be useful for clinical purposes but are not used for scoring.

Response options. Of the 19 items included in scoring, items 1–4 have free-entry responses asking for usual bedtime and wake up times, number of minutes to fall asleep, and hours slept per night. Items 5–17 have 4-point Likert scale responses relating to frequency of specified sleep problems. Item 18 has a 4-point Likert scale response relating to overall assessment of sleep quality ("very good," "fairly good," "fairly bad," or "very bad"). Item 19 has a 4-point Likert response scale relating to the respondent's overall assessment of "enthusiasm to get things done" ("no problem at all," "only a very slight problem," "somewhat of a problem," or "a very big problem").

Recall period for items. Last month.

Endorsements. No.

Examples of use. In multiple disease areas, the PSQI has often been used as an outcome in clinical trials of interventions intended to reduce sleep disturbances (75–81). It has been used in clinical trials to define inclusion criteria for poor sleep quality (e.g., participants with PSQI scores >5 were eligible for inclusion) (82). The PSQI has also been used to determine the impact of a particular sleep disturbance, such as nocturnal hypoxemia in chronic obstructive pulmonary disease, on sleep quality (44). The PSQI has been used as an outcome in epidemiologic studies intending to determine risk factors for, or prevalence of, poor sleep quality in various populations, including those with rheumatoid arthritis, chronic pain, fibromyalgia, and chronic opiate usage (22,83–86).

Practical Application

How to obtain. Questionnaire and scoring instructions are available in the appendix of the original validating publication (74). Permission for use can be obtained from the author, Daniel J. Buysse, MD, University of Pittsburgh, 3811 O'Hara Street, E-1127, Pittsburgh, PA 15213. E-mail: buyssedj@upmc.edu.

Method of administration. Self-administered written questionnaire.

Scoring. Each of the 7 component scores is determined based on scoring algorithms, with the 7 component scores each yielding a score of 0–3. A PSQI global (total) score is obtained by summing each of the 7 component scores. Scoring algorithms for each component involve a mixture of averaging Likert response scores, categorization of free-text responses (e.g., sleep latency of 15–30 minutes = 1 point), and arithmetic determination of sleep efficiency based on free-text responses.

Score interpretation. Score range is 0–21 points, with higher scores indicating better sleep quality. In the original validation report, a PSQI global score >5 correctly identified 88.5% as "good sleepers" versus "poor sleepers," with a sensitivity of 89.6% and a specificity of 86.5% (74). However, accuracy has been less high in other populations: 1) a threshold score of 5 was 72% sensitive and 55% specific among Nigerian university students (87), and 2) in a heterogeneous population (most with history of malignancy or renal transplant), a threshold score of 8 appeared more appropriate (88). Among Chinese-speaking patients, a PSQI score >5 was 98% sensitive and 55% specific for insomnia (89).

Respondent burden. Time to complete is reported to be 5–10 minutes (74).

Administration burden. Time to score is reported to be 5 minutes (74). Because of the need to integrate various responses and calculate such variables as sleep efficiency, hand-calculation of scores may be somewhat burdensome, but a scoring algorithm can readily be incorporated into statistical programming software or a spreadsheet for automated calculation.

Translations/adaptations. Validated versions of the PSQI are available in Spanish, French, Japanese, Chinese, Greek, German, Hebrew, Persian, and Arabic (89–98).

Psychometric Information

Method of development. The PSQI was derived from "clinical intuition and experience with sleep disorder patients; a review of previous sleep quality questionnaires reported in the literature; and clinical experience with the instrument during 18 months of field testing"(74).

Acceptability. Total scores appear reasonably normal in distribution in both healthy populations and in those with higher frequency of sleep disturbances (74). Buysse et al report that 6.3% of 158 respondents failed to give complete responses to all items and scores could not therefore be calculated. In a validating study among cancer patients, PSQI scores for 21% of respondents could not be calculated due to missing responses. The presence of free-text items is associated with greater nonresponse; the plurality of missing items reported by Beck et al (99) was due to missing free-text responses necessary to calculate sleep efficiency. Interviewer followup after completion of the questionnaire to query about missing items reduced the percentage of scores that could not be calculated to 4.2%.

Reliability. In the original validating study, the 7 component scores of the PSQI had an overall Cronbach's alpha of $\alpha = 0.83$, and individual items were strongly correlated with one another, also with $\alpha = 0.83$ (74). In separate studies with different populations, the Cronbach's alpha scores have been similar (88,99). Test–retest reliability (Pearson's correlation coefficient) for the global PSQI was 0.85 ($P < 0.001$) when testing was separated by ~4 weeks (74). Among German-speaking respondents with insomnia, the test–retest Pearson's correlation coefficients were 0.90 and 0.86, based on testing separated in time by 2 days and mean 45.6 days, respectively (97).

Validity. *Criterion validity.* Based on the gold standard of clinical evaluation, the PSQI distinguished "good sleepers" from "poor sleepers" with reasonable accuracy in its original validation, which was a chief basis for demonstrating initial validity (see Score Interpretation section above) (74).

Concurrent validity. In the original validation, the sleep latency component of the PSQI was modestly correlated with sleep latency on single-night polysomnography (PSG) ($r = 0.33$, $P < 0.001$), and global PSQI scores were also weakly correlated with PSG sleep latency ($r = 0.20$, $P < 0.01$). Other correlations with PSG results were, for the most part, not significant (74), and in a recent study, Buysse et al concluded that the PSQI is not likely be useful as a screening measure for PSG sleep abnormalities (30). A variety of other studies have demonstrated PSQI concurrent validity: 1) PSQI component scores were correlated with sleep duration ($r = 0.81$) and sleep latency ($r = 0.71$) as assessed by daily sleep diaries among insomnia patients (97), 2) PSQI global scores were correlated with Insomnia Severity Index ($r = 0.76$) among Arabic-speaking patients (96), and 3) PSQI global scores were correlated with sleep-related items from the Symptoms Experience Report and with sleep-related items from the Centers for Epidemiological Studies Depression Scale (88).

Factor validity. Based on the original formulation of the PSQI as a measure of sleep quality, Buysse et al suggested that its 7 components be combined into a single factor, the PSQI global score (74). However, in a factor analysis later conducted by Cole et al (including Daniel Buysse, lead author of the original validation study), a 3-factor scoring model provided significantly better fit than the original single-factor model, where the 3 factors are sleep efficiency, perceived sleep quality, and daily disturbances (100). Such a scoring model has not thus far been widely accepted and has not yet been further validated.

Ability to detect change. The PSQI has demonstrated sensitivity to change by virtue of clinical trial interventions intended to reduce sleep disturbances, which have shown an improvement in PSQI scores, along with concomitant improvement in other sleep-related measures (75–80).

Discussion

The PSQI is a widely used measure of sleep quality that is more global in nature than other measures reviewed here. The PSQI includes elements of daytime dysfunction, captured more specifically in the FOSQ. Three of the 7 PSQI components (sleep latency, sleep duration, and sleep efficiency) are often elicited to identify evidence of insomnia (101). However, unlike the ISI, these 3 components are based largely on free-text numerical responses that are used to quantify these components, whereas the ISI asks, with Likert responses, about perceived respondent difficulties related to these components. The PSQI also includes 1 item inquiring about daytime sleepiness, although Buysse has argued that the PSQI and Epworth Sleepiness Scale correlate weakly with each other (r = 0.16) and measure orthogonal dimensions of sleep-wake symptoms (30). One strength of the PSQI is, therefore, the broad range of its coverage in measuring several aspects of sleep quality and combining these into a global score. One drawback is the potential disagreement about whether the PSQI represents a single factor (100).

AUTHOR CONTRIBUTIONS

Dr. Omachi drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Ad Hoc Committee on Systemic Lupus Erythematosus Response Criteria for Fatigue. Measurement of fatigue in systemic lupus erythematosus: a systematic review. Arthritis Rheum 2007;57:1348–57.
2. Stebbings S, Herbison P, Doyle TC, Treharne GJ, Highton J. A comparison of fatigue correlates in rheumatoid arthritis and osteoarthritis: disparity in associations with disability, anxiety and sleep disturbance. Rheumatology (Oxford) 2010;49:361–7.
3. Abad VC, Sarinas PS, Guilleminault C. Sleep and rheumatologic disorders. Sleep Med Rev 2008;12:211–28.
4. Drewes AM. Pain and sleep disturbances with special reference to fibromyalgia and rheumatoid arthritis. Rheumatology (Oxford) 1999; 38:1035–8.
5. Chandrasekhara PK, Jayachandran NV, Rajasekhar L, Thomas J, Narsimulu G. The prevalence and associations of sleep disturbances in patients with systemic lupus erythematosus. Mod Rheumatol 2009; 19:407–15.
6. Gudbjornsson B, Hetta J. Sleep disturbances in patients with systemic lupus erythematosus: a questionnaire-based study. Clin Exp Rheumatol 2001;19:509–14.
7. Costa DD, Bernatsky S, Dritsa M, Clarke AE, Dasgupta K, Keshani A, et al. Determinants of sleep quality in women with systemic lupus erythematosus. Arthritis Rheum 2005;53:272–8.
8. Wolfe F, Michaud K, Li T. Sleep disturbance in patients with rheumatoid arthritis: evaluation by medical outcomes study and visual analog sleep scales. J Rheumatol 2006;33:1942–51.
9. Taylor-Gjevre RM, Gjevre JA, Skomro R, Nair B. Restless legs syndrome in a rheumatoid arthritis patient cohort. J Clin Rheumatol 2009;15:12–5.
10. Reading SR, Crowson CS, Rodeheffer RJ, Fitz-Gibbon PD, Maradit-Kremers H, Gabriel SE. Do rheumatoid arthritis patients have a higher risk for sleep apnea? J Rheumatol 2009;36:1869–72.
11. Turner GA, Lower EE, Corser BC, Gunther KL, Baughman RP. Sleep apnea in sarcoidosis. Sarcoidosis Vasc Diffuse Lung Dis 1997;14:61–4.
12. Iaboni A, Ibanez D, Gladman DD, Urowitz MB, Moldofsky H. Fatigue in systemic lupus erythematosus: contributions of disordered sleep, sleepiness, and depression. J Rheumatol 2006;33:2453–7.
13. Lee YC, Chibnik LB, Lu B, Wasan AD, Edwards RR, Fossel AH, et al. The relationship between disease activity, sleep, psychiatric distress and pain sensitivity in rheumatoid arthritis: a cross-sectional study. Arthritis Res Ther 2009;11:R160.
14. Majde JA, Krueger JM. Links between the innate immune system and sleep. J Allergy Clin Immunol 2005;116:1188–98.
15. Luyster FS, Chasens ER, Wasko MC, Dunbar-Jacob J. Sleep quality and functional disability in patients with rheumatoid arthritis. J Clin Sleep Med 2011;7:49–55.
16. Johns MW. A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. Sleep 1991;14:540–5.
17. Hogl B, Saletu M, Brandauer E, Glatzl S, Frauscher B, Seppi K, et al. Modafinil for the treatment of daytime sleepiness in Parkinson's disease: a double-blind, randomized, crossover, placebo-controlled polygraphic trial. Sleep 2002;25:905–9.
18. Rammohan KW, Rosenberg JH, Lynn DJ, Blumenfeld AM, Pollak CP, Nagaraja HN. Efficacy and safety of modafinil (Provigil) for the treatment of fatigue in multiple sclerosis: a two centre phase 2 study. J Neurol Neurosurg Psychiatry 2002;72:179–83.
19. Teodorescu M, Consens FB, Bria WF, Coffey MJ, McMorris MS, Weatherwax KJ, et al. Correlates of daytime sleepiness in patients with asthma. Sleep Med 2006;7:607–13.
20. Wang R, Zou D, Ma X, Zhao Y, Yan X, Yan H, et al. Impact of gastroesophageal reflux disease on daily life: the Systematic Investigation of Gastrointestinal Diseases in China (SILC) epidemiological study. Health Qual Life Outcomes 2010;8:128.
21. Alvarez Lario B, Alonso Valdivielso JL, Alegre Lopez J, Martel Soteres C, Viejo Banuelos JL, Maranon Cabello A. Fibromyalgia syndrome: overnight falls in arterial oxygen saturation. Am J Med 1996;101:54–60.
22. Menefee LA, Frank ED, Doghramji K, Picarello K, Park JJ, Jalali S, et al. Self-reported sleep quality and quality of life for individuals with chronic pain conditions. Clin J Pain 2000;16:290–7.
23. Johns M, Hocking B. Daytime sleepiness and sleep habits of Australian workers. Sleep 1997;20:844–9.
24. Chen NH, Johns MW, Li HY, Chu CC, Liang SC, Shu YH, et al. Validation of a Chinese version of the Epworth Sleepiness Scale. Qual Life Res 2002;11:817–21.
25. Chiner E, Arriero JM, Signes-Costa J, Marco J, Fuentes I. Validation of the Spanish version of the Epworth Sleepiness Scale in patients with a sleep apnea syndrome. Arch Bronconeumol 1999;35:422–7. In Spanish.
26. Izci B, Ardic S, Firat H, Sahin A, Altinors M, Karacan I. Reliability and validity studies of the Turkish version of the Epworth Sleepiness Scale. Sleep Breath 2008;12:161–8.
27. Tsara V, Serasli E, Amfilochiou A, Constantinidis T, Christaki P. Greek version of the Epworth Sleepiness Scale. Sleep Breath 2004;8: 91–5.
28. Bloch KE, Schoch OD, Zhang JN, Russi EW. German version of the Epworth Sleepiness Scale. Respiration 1999;66:440–7.
29. Schmidt-Nowara WW, Wiggins CL, Walsh JK, Bauer C. Prevalence of sleepiness in an adult population. Sleep Res 1989;18:302.
30. Buysse DJ, Hall ML, Strollo PJ, Kamarck TW, Owens J, Lee L, et al. Relationships between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and clinical/polysomnographic measures in a community sample. J Clin Sleep Med 2008;4:563–71.
31. Johns MW. Reliability and factor analysis of the Epworth Sleepiness Scale. Sleep 1992;15:376–81.
32. Johns MW. Sleepiness in different situations measured by the Epworth Sleepiness Scale. Sleep 1994;17:703–10.
33. Nguyen AT, Baltzan MA, Small D, Wolkove N, Guillon S, Palayew M. Clinical reproducibility of the Epworth Sleepiness Scale. J Clin Sleep Med 2006;2:170–4.
34. Chervin RD, Aldrich MS, Pickett R, Guilleminault C. Comparison of the results of the Epworth Sleepiness Scale and the Multiple Sleep Latency Test. J Psychosom Res 1997;42:145–55.
35. Olson LG, Cole MF, Ambrogetti A. Correlations among Epworth Sleepiness Scale scores, multiple sleep latency tests and psychological symptoms. J Sleep Res 1998;7:248–53.
36. Johns MW. Sensitivity and specificity of the Multiple Sleep Latency Test (MSLT), the Maintenance of Wakefulness Test and the Epworth Sleepiness Scale: failure of the MSLT as a gold standard. J Sleep Res 2000;9:5–11.

37. Gay PC, Herold DL, Olson EJ. A randomized, double-blind clinical trial comparing continuous positive airway pressure with a novel bilevel pressure system for treatment of obstructive sleep apnea syndrome. Sleep 2003;26:864–9.
38. Fatigue Guidelines Development Panel of the Multiple Sclerosis Council for Clinical Practice Guidelines. Fatigue and multiple sclerosis: evidence based management strategies for fatigue in multiple sclerosis. Washington (DC): Multiple Sclerosis Council; 1998.
39. Weaver TE, Laizner AM, Evans LK, Maislin G, Chugh DK, Lyon K, et al. An instrument to measure functional status outcomes for disorders of excessive sleepiness. Sleep 1997;20:835–43.
40. Chasens ER, Ratcliffe SJ, Weaver TE. Development of the FOSQ-10: a short version of the Functional Outcomes of Sleep Questionnaire. Sleep 2009;32:915–9.
41. Hughes K, Glass C, Ripchinski M, Gurevich F, Weaver TE, Lehman E, et al. Efficacy of the topical nasal steroid budesonide on improving sleep and daytime somnolence in patients with perennial allergic rhinitis. Allergy 2003;58:380–5.
42. Weaver TE, Chasens ER, Arora S. Modafinil improves functional outcomes in patients with residual excessive sleepiness associated with CPAP treatment. J Clin Sleep Med 2009;5:499–505.
43. Walker RP, Paloyan E, Gopalsami C. Symptoms in patients with primary hyperparathyroidism: muscle weakness or sleepiness. Endocr Pract 2004;10:404–8.
44. Lewis CA, Fergusson W, Eaton T, Zeng I, Kolbe J. Isolated nocturnal desaturation in COPD: prevalence and impact on quality of life and sleep. Thorax 2009;64:133–8.
45. Burke CK, Peirce JM, Kidorf MS, Neubauer D, Punjabi NM, Stoller KB, et al. Sleep problems reported by patients entering opioid agonist treatment. J Subst Abuse Treat 2008;35:328–33.
46. Carmona-Bernal C, Ruiz-Garcia A, Villa-Gil M, Sanchez-Armengol A, Quintana-Gallego E, Ortega-Ruiz F, et al. Quality of life in patients with congestive heart failure and central sleep apnea. Sleep Med 2008;9:646–51.
47. Shaheen NJ, Madanick RD, Alattar M, Morgan DR, Davis PH, Galanko JA, et al. Gastroesophageal reflux disease as an etiology of sleep disturbance in subjects with insomnia and minimal reflux symptoms: a pilot study of prevalence and response to therapy. Dig Dis Sci 2008;53:1493–9.
48. Teixeira VG, Faccenda JF, Douglas NJ. Functional status in patients with narcolepsy. Sleep Med 2004;5:477–83.
49. Dellaripa PF, Fry TA, Willoughby J, Arndt WF, Angelakis WJ, Campagna AC. The treatment of interstitial lung disease associated with rheumatoid arthritis with infliximab [abstract]. Chest 2003;124:109S
50. Mermigkis C, Stagaki E, Amfilochiou A, Polychronopoulos V, Korkonikitas P, Mermigkis D, et al. Sleep quality and associated daytime consequences in patients with idiopathic pulmonary fibrosis. Med Princ Pract 2009;18:10–5.
51. Ferrer M, Vilagut G, Monasterio C, Montserrat JM, Mayos M, Alonso J. Measurement of the perceived impact of sleep problems: the Spanish version of the functional outcomes sleep questionnaire and the Epworth Sleepiness Scale. Med Clin (Barc) 1999;113:250–5. In Spanish.
52. Izci B, Firat H, Ardic S, Kokturk O, Gelir E, Altinors M. Adaptation of Functional Outcomes of Sleep Questionnaire (FOSQ) to Turkish population. Tuberk Toraks 2004;52:224–30.
53. Stavem K, Kjelsberg FN, Ruud EA. Reliability and validity of the Norwegian version of the Functional Outcomes of Sleep Questionnaire. Qual Life Res 2004;13:541–9.
54. Vidal S, Ferrer M, Masuet C, Somoza M, Martinez Ballarin JI, Monasterio C. Spanish version of the Functional Outcomes of Sleep Questionnaire: scores of healthy individuals and of patients with sleep apnea-hypopnea syndrome. Arch Bronconeumol 2007;43:256–61. In Spanish.
55. Buttner A, Feier C, Galetke W, Ruhle K. A questionnaire to capture the functional effects of daytime drowsiness on quality of life in case of obstructive sleep apnea syndrome: Functional Outcomes of Sleep Questionnaire (FOSQ). Pneumologie 2008;62:548–52. In German.
56. Gooneratne NS, Weaver TE, Cater JR, Pack FM, Arner HM, Greenberg AS, et al. Functional outcomes of excessive daytime sleepiness in older adults. J Am Geriatr Soc 2003;51:642–9.
57. Moyer CA, Sonnad SS, Garetz SL, Helman JI, Chervin RD. Quality of life in obstructive sleep apnea: a systematic review of the literature. Sleep Med 2001;2:477–91.
58. Belanger L, Morin CM, Langlois F, Ladouceur R. Insomnia and generalized anxiety disorder: effects of cognitive behavior therapy for gad on insomnia symptoms. J Anxiety Disord 2004;18:561–71.
59. Joffe H, Petrillo L, Viguera A, Koukopoulos A, Silver-Heilman K, Farrell A, et al. Eszopiclone improves insomnia and depressive and anxious symptoms in perimenopausal and postmenopausal women with hot flashes: a randomized, double-blinded, placebo-controlled crossover trial. Am J Obstet Gynecol 2010;202:171.e1–11.
60. Riley WT, Mihm P, Behar A, Morin CM. A computer device to deliver behavioral interventions for insomnia. Behav Sleep Med 2010;8:2–15.
61. Roth T, Price JM, Amato DA, Rubens RP, Roach JM, Schnitzer TJ. The effect of eszopiclone in patients with insomnia and coexisting rheumatoid arthritis: a pilot study. Prim Care Companion J Clin Psychiatry 2009;11:292–301.
62. Savard J, Simard S, Ivers H, Morin CM. Randomized study on the efficacy of cognitive-behavioral therapy for insomnia secondary to breast cancer, part I: sleep and psychological effects. J Clin Oncol 2005;23:6083–96.
63. Tang NK, Wright KJ, Salkovskis PM. Prevalence and correlates of clinical insomnia co-occurring with chronic back pain. J Sleep Res 2007;16:85–95.
64. Walsh JK, Krystal AD, Amato DA, Rubens R, Caron J, Wessel TC, et al. Nightly treatment of primary insomnia with eszopiclone for six months: effect on sleep, quality of life, and work limitations. Sleep 2007;30:959–68.
65. Daley M, Morin CM, LeBlanc M, Gregoire JP, Savard J, Baillargeon L. Insomnia and its relationship to health-care utilization, work absenteeism, productivity and accidents. Sleep Med 2009;10:427–38.
66. Viitanen J, Ronni S, Ala-Peijari S, Uoti-Reilama K, Kautiainen H. A comparison of self-estimated symptoms and impact of disease in fibromyalgia and rheumatoid arthritis. J Musculoskelet Pain 2000;8:21–33.
67. Bastien CH, Vallieres A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med 2001;2:297–307.
68. Savard MH, Savard J, Simard S, Ivers H. Empirical validation of the Insomnia Severity Index in cancer patients. Psychooncology 2005;14:429–41.
69. Sierra JC, Guillen-Serrano V, Santos-Iglesias P. Insomnia Severity Index: some indicators about its reliability and validity on an older adults sample. Rev Neurol 2008;47:566–70. In Spanish.
70. Yu DS. Insomnia Severity Index: psychometric properties with Chinese community-dwelling older people. J Adv Nurs 2010;66:2350–9.
71. Morin CM. Insomnia: psychological assessment and management. New York: Guilford Press; 1993.
72. Morin CM, Colecchi C, Stone J, Sood R, Brink D. Behavioral and pharmacological therapies for late-life insomnia: a randomized controlled trial. JAMA 1999;281:991–9.
73. Yang M, Morin CM, Schaefer K, Wallenstein GV. Interpreting score differences in the Insomnia Severity Index: using health-related outcomes to define the minimally important difference. Curr Med Res Opin 2009;25:2487–94.
74. Buysse DJ, Reynolds CF 3rd, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28:193–213.
75. Berger AM, Kuhn BR, Farr LA, Von Essen SG, Chamberlain J, Lynch JC, et al. One-year outcomes of a behavioral therapy intervention trial on sleep quality and cancer-related fatigue. J Clin Oncol 2009;27:6033–40.
76. Inoue Y, Kuroda K, Hirata K, Uchimura N, Kagimura T, Shimizu T. Long-term open-label study of pramipexole in patients with primary restless legs syndrome. J Neurol Sci 2010;294:62–6.
77. Irwin MR, Olmstead R, Motivala SJ. Improving sleep quality in older adults with moderate sleep complaints: a randomized controlled trial of Tai Chi Chih. Sleep 2008;31:1001–8.
78. Johnson DA, Orr WC, Crawley JA, Traxler B, McCullough J, Brown KA, et al. Effect of esomeprazole on nighttime heartburn and sleep quality in patients with GERD: a randomized, placebo-controlled trial. Am J Gastroenterol 2005;100:1914–22.
79. Reid KJ, Baron KG, Lu B, Naylor E, Wolfe L, Zee PC. Aerobic exercise improves self-reported sleep and quality of life in older adults with insomnia. Sleep Med 2010;11:934–40.
80. Rondanelli M, Opizzi A, Monteferrario F, Antoniello N, Manni R, Klersy C. The effect of melatonin, magnesium, and zinc on primary insomnia in long-term care facility residents in Italy: a double-blind, placebo-controlled clinical trial. J Am Geriatr Soc 2011;59:82–90.
81. Skomro RP, Gjevre J, Reid J, McNab B, Ghosh S, Stiles M, et al. Outcomes of home-based diagnosis and treatment of obstructive sleep apnea. Chest 2010;138:257–63.
82. Cunningham JM, Blake C, Power CK, O'Keeffe D, Kelly V, Horan S, et al. The impact on sleep of a multidisciplinary cognitive behavioural pain management programme: a pilot study. BMC Musculoskelet Disord 2011;12:5.
83. Cakirbay H, Bilici M, Kavakci O, Cebi A, Guler M, Tan U. Sleep quality and immune functions in rheumatoid arthritis patients with and without major depression. Int J Neurosci 2004;114:245–56.
84. Marin R, Cyhan T, Miklos W. Sleep disturbance in patients with chronic low back pain. Am J Phys Med Rehabil 2006;85:430–5.
85. Osorio CD, Gallinaro AL, Lorenzi-Filho G, Lage LV. Sleep quality in patients with fibromyalgia using the Pittsburgh Sleep Quality Index. J Rheumatol 2006;33:1863–5.

86. Stein MD, Herman DS, Bishop S, Lassor JA, Weinstock M, Anthony J, et al. Sleep disturbances among methadone maintained patients. J Subst Abuse Treat 2004;26:175–80.
87. Aloba OO, Adewuya AO, Ola BA, Mapayi BM. Validity of the Pittsburgh Sleep Quality Index (PSQI) among Nigerian university students. Sleep Med 2007;8:266–70.
88. Carpenter JS, Andrykowski MA. Psychometric evaluation of the Pittsburgh Sleep Quality Index. J Psychosom Res 1998;45:5–13.
89. Tsai PS, Wang SY, Wang MY, Su CT, Yang TT, Huang CJ, et al. Psychometric evaluation of the Chinese version of the Pittsburgh Sleep Quality Index (CPSQI) in primary insomnia and control subjects. Qual Life Res 2005;14:1943–52.
90. Blais FC, Gendron L, Mimeault V, Morin CM. Evaluation of insomnia: validity of 3 questionnaires. Encephale 1997;23:447–53. In French.
91. Doi Y, Minowa M, Uchiyama M, Okawa M, Kim K, Shibui K, et al. Psychometric assessment of subjective sleep quality using the Japanese version of the Pittsburgh Sleep Quality Index (PSQI-J) in psychiatric disordered and control subjects. Psychiatry Res 2000;97:165–72.
92. Farrahi J, Nakhaee N, Sheibani V, Garrusi B, Amirka. A. Psychometric properties of the Persian version of the Pittsburgh Sleep Quality Index addendum for PTSD (PSQI-A). Sleep Breath 2009;13:259–62.
93. Jimenez-Genchi A, Monteverde-Maldonado E, Nenclares-Portocarrero A, Esquivel-Adame G, de la Vega-Pacheco A. Reliability and factorial analysis of the Spanish version of the Pittsburg Sleep Quality Index among psychiatric patients. Gac Med Mex 2008;144:491–6. In Spanish.
94. Kotronoulas GC, Papadopoulou CN, Papapetrou A, Patiraki E. Psychometric evaluation and feasibility of the Greek Pittsburgh Sleep Quality Index (GR-PSQI) in patients with cancer receiving chemotherapy. Support Care Cancer 2010. E-pub ahead of print.
95. Shochat T, Tzischinsky O, Oksenberg A, Peled R. Validation of the Pittsburgh Sleep Quality Index Hebrew translation (PSQI-H) in a sleep clinic sample. Isr Med Assoc J 2007;9:853–6.
96. Suleiman KH, Yates BC, Berger AM, Pozehl B, Meza J. Translating the Pittsburgh Sleep Quality Index into Arabic. West J Nurs Res 2010;32: 250–68.
97. Backhaus J, Junghanns K, Broocks A, Riemann D, Hohagen F. Test-retest reliability and validity of the Pittsburgh Sleep Quality Index in primary insomnia. J Psychosom Res 2002;53:737–40.
98. Simeit R, Deck R, Conta-Marx B. Sleep management training for cancer patients with insomnia. Support Care Cancer 2004;12:176–83.
99. Beck SL, Schwartz AL, Towsley G, Dudley W, Barsevick A. Psychometric evaluation of the Pittsburgh Sleep Quality Index in cancer patients. J Pain Symptom Manage 2004;27:140–8.
100. Cole JC, Motivala SJ, Buysse DJ, Oxman MN, Levin MJ, Irwin MR. Validation of a 3-factor scoring model for the Pittsburgh sleep quality index in older adults. Sleep 2006;29:112–6.
101. Roth T. Insomnia: definition, prevalence, etiology, and consequences. J Clin Sleep Med 2007;3 Suppl:S7–10.

Table 1. Summary table for sleep measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
ESS	Measures sleepiness as likelihood of falling asleep in various situations	Written	2–3 minutes	<1 minute	Range 0–24 ≥11 is positive for EDS	α = 0.74–0.88† Test–retest reliability 0.82 after 5 months	Concurrent validity based on correlation with MSLT and ability to predict narcolepsy diagnoses	Sensitive to change in clinical trials MCID not reported	Short Widely used Simple concept	Cannot discern if sleepiness is due to sleep disturbance or other causes, e.g., medications
FOSQ	Measure functional impairment, resulting from sleepiness, in ADLs and recreational activities	Written	15 minutes for FOSQ-30	3–5 minutes	Range 5–20 Higher scores = better functional status	α = 0.95† Test–retest reliability 0.90 after 1 week	Concurrent validity based on correlations with SIP and SF-36 subscales Discriminant validity-based to classify respondents with sleep disorders	Sensitive to change in clinical trials MCID not reported	Widely used Measures HRQoL related to sleepiness but not specific to any disease	Not widely applied in rheumatologic diseases FOSQ-10 is recently introduced shorter version but with limited application so far Questions about sexual function associated with higher nonresponse
ISI	Measure severity of insomnia symptoms as difficulty initiating and maintaining sleep and as perceived consequences of insomnia	Written or clinician-administered	<5 minutes	<1 minute	Range: 0–28 Higher scores indicate greater insomnia symptoms Suggested but not validated guidelines‡	α = 0.76–0.90† Test–retest reliability 0.83 after 1 month	Concurrent validity based primarily on correlations with sleep diary	Sensitive to change in clinical trials MCID proposed as 6 points = 1.5 SDs	Short High face validity based on similarity to DSM-IV criteria for insomnia Widely used	Does not elucidate cause of insomnia, whether related to psychological factors, pain, or other symptoms
PSQI	Measure overall sleep quality across multiple dimensions, including insomnia symptoms, functional impairment, sleepiness, and causes of sleep disturbances	Written	5–10 minutes	5 minutes	Range 0–21 Scores >5 indicate poor sleep quality	α = 0.83† Test–retest reliability 0.85 after 4 weeks	Criterion validity based on 88.5% accuracy in identifying good sleepers vs. poor sleepers Concurrent validity-based correlation with certain PSG variables, sleep diary, and other sleep-related instruments	Sensitive to change in clinicaltrials MCID notreported	Broad measure of sleep quality capturing multiple dimensions Widely used	Potential disagreement about whether PSQI represents a single factor Free-text responses associated with higher nonresponse unless interviewer followup enacted

* ESS = Epworth Sleepiness Scale; EDS = extreme daytime sleepiness; MSLT = multiple sleep latency tests; MCID = minimal clinically important difference; FOSQ = Functional Outcome of Sleep Questionnaire; ADL = activity of daily living; SIP = Sickness Impact Profile; SF-36 = Short Form 36 Health Survey; HRQOL = health-related quality of life; ISI = Insomnia Severity Index; DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; PSQI = Pittsburgh Sleep Quality Index; PSG = polysomnography.

† Cronbach's alpha.

‡ Guidelines: 0–7 no insomnia; 8–14 subthreshold insomnia; 15–21 clinical insomnia, 22–28 severe clinical insomnia.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S189–S199
DOI 10.1002/acr.20631
© 2011, American College of Rheumatology

Measures of Hand Function

Arthritis Hand Function Test (AHFT), Australian Canadian Osteoarthritis Hand Index (AUSCAN), Cochin Hand Function Scale, Functional Index for Hand Osteoarthritis (FIHOA), Grip Ability Test (GAT), Jebsen Hand Function Test (JHFT), and Michigan Hand Outcomes Questionnaire (MHQ)

JANET L. POOLE

INTRODUCTION

Many of the rheumatic diseases result in pain, deformities, weakness, and other impairments that affect the hands. However, measures of these impairments may not provide information about the ability to use the hands for self-care, work, and leisure activities. Many health status, quality of life, and functional ability assessments have several questions pertaining to hand function, but these are usually limited to a few items due to the large scope of the assessment. For some individuals, specific assessments of hand function may be warranted to measure hand function and document treatment effectiveness. This section will review self-report and performance-based hand function tests that have been used with persons with rheumatic diseases and have psychometric support.

ARTHRITIS HAND FUNCTION TEST (AHFT)

Description

Purpose. The AHFT is an 11-item performance-based test designed to measure hand strength and dexterity in persons with arthritis.

Content. The items include grip and pinch strength, pegboard dexterity, lacing a shoe and tying a bow, fastening/unfastening 4 buttons, fastening/unfastening 2 safety pins, cutting putty with a knife and fork, manipulating coins into a slot, lifting a tray of tin cans, and pouring a glass of water. For the grip, pinch, and pegboard dexterity, each hand is tested separately.

Number of items. 4 subscales: grip and pinch strength (3 items), dexterity (1 item), applied dexterity (5 items), and applied strength (2 items).

Response options/scale. Grip strength is measured in mm Hg, while pinch strength is measured in kg. Dexterity and applied dexterity items are timed in seconds. Applied strength is the number of cans lifted and volume of water lifted in the pitcher in ml.

Recall period for items. Not applicable (N/A).

Endorsements. N/A.

Examples of use. See reference list.

Practical Application

How to obtain. Contact Catherine Backman and Hazel Mackie at School of Rehabilitation Sciences, University of British Columbia, T325-2211 Wesbrook Mall, Vancouver, British Columbia V6T2B5, Canada.

Cost. Pinchmeter and theraplast must be ordered from an adapted equipment catalog for ~$300. Other equipment such as the pegboard, coin box, putty guide, and tray must be constructed. Other equipment is easily available from a discount store. A cheap suitcase is helpful to transport items. In total, the entire cost according to one of the authors is ~$500.00.

Method of administration. Performance-based test; test manual describes administration setup and instructions.

Scoring. The test manual purchased from the authors describes specific scoring instructions. Computer scoring is not necessary. There are no instructions for handling missing values.

Score interpretation. Score range: grip strength range 0–300 mm Hg, pinch strength range 0–30 kg, applied strength range for cans 0–12 cans, applied strength range for pouring water 0–2,000 ml, and dexterity and applied dexterity range 0 to undetermined number of seconds.

Scores can be transferred to a hand function profile sheet matched for age and sex. This profile provides a summary and compares the scores to the norms. Normative values are available in the manual.

Janet L. Poole, PhD, OTR/L: University of New Mexico, Albuquerque.

Address correspondence to Janet L. Poole, PhD, OTR/L, Occupational Therapy Graduate Program, Health Sciences and Services Building, Room 215, 1 University of New Mexico, MSC90, Albuquerque, NM 87131-0011. E-mail: jpoole@salud.unm.edu.

Submitted for publication February 11, 2011; accepted in revised form June 7, 2011.

Respondent burden. Time to complete is 20–30 minutes, depending on skill and administrators' familiarity with items. Cutting the putty can be difficult.

Administrative burden. Time to administer is 20–30 minutes, depending on the skill and familiarity of items of the administrator. Time to score is 5 minutes; items are scored as the test is administered. Training is necessary as the administrator has to be familiar with setup and administration of items.

Translations/adaptations. Languages available: English. Cultural adaptations: N/A.

Psychometric Information

Method of development. Content or face validity: items were developed based on a systemic review of other hand function tests. Items were reviewed by 5 occupational therapists who judged the final items to be clear and important unilateral and bilateral tasks. Patients were not involved in development of the test and item-response theory was not used in development or item selection.

Acceptability. Missing data are common. Floor or ceiling effects are possible.

Reliability. *Interrater reliability: rheumatoid arthritis (RA).* Intraclass correlation coefficient (ICC) ranged from 0.89–1.0 between 2 independent observers rating 20 subjects (1).

Interrater reliability: osteoarthritis (OA). ICC ranged from 0.99–1.0 between 2 independent observers rating 26 subjects (2).

Interrater reliability: RA and OA. Pearson's correlations ranged from 0.45–0.99 between 6 self-trained raters assessing 30 subjects (3).

Interrater reliability: systemic sclerosis (SSc). ICC ranged from 0.99–1.0 between 2 independent observers rating 20 subjects (4).

Test–retest reliability: RA. Twenty subjects were tested twice within 2 weeks. ICC ranged from 0.53–0.96 (1).

Test–retest reliability: OA. Twenty-six subjects were tested twice within 2 weeks. ICC ranged from 0.7–0.96 (2).

Test–retest reliability: SSc. Twenty subjects were tested twice within 2 weeks. ICC ranged from 0.80–0.97 (5).

Validity. *Concurrent validity: healthy controls.* Three hundred ninety-five healthy adult volunteers were administered the 9-hole pegboard test and items from the applied dexterity section of the AHFT. Correlations ranged from 0.32–0.60 (5).

Convergent validity: RA. Twenty subjects were evaluated with the AHFT and Jebsen Hand Function Test (JHFT). Correlations between the AHFT and scores on the JHFT were 0.61–0.64 for the right hand scores and 0.02–0.08 for the left hand scores (1). ICC was 0.71 between scores on the AHFT and the dexterity subscale of the Arthritis Impact Measurement Scales (AIMS) (1).

Convergent validity: OA. Twenty-six subjects were evaluated with the AHFT and self-reports of physical activities of daily living (PADL) and instrumental activities of daily living (IADL). Correlations between the AHFT and PADL scores ranged from 0.40–0.69 and between the AHFT and IADL scores ranged from 0.46–0.75 (2). For 40 subjects with hand OA, correlations between the AHFT and self-reports of hand function, including the Michigan Hand Questionnaire (rs = 0.3–0.65), Cochin Hand Function Disability Scale (rs = 0.52–0.64), and Dreiser's Functional Index for Hand OA (rs = 0.44–0.57), were fair to good, with stronger correlation for the strength items (6).

Convergent validity: SSc. Twenty subjects with SSc were evaluated with the AHFT, Health Assessment Questionnaire (HAQ), and physical component of the AIMS2. Correlations ranged from 0.32–0.73 with the HAQ and from 0.19–0.69 with the AIMS2 (4).

Ability to detect change. Studies have not been done.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The majority of hand function tests assess only 1 aspect of function, such as strength or dexterity, and only unilateral tasks, and do not include functional tasks. The AHFT is a performance-based test, which measures both unilateral and bilateral functional tasks.

The test has adequate psychometric properties for RA, OA, and SSc. Predictive validity and responsiveness to change have not been documented. There is no summative total score, as most of the items are measured in different units of measure, which is a disadvantage when using the AHFT in research, as numerous correlations or comparisons must be made for each of the 11 items and not just 1 score. Groups or conditions for which the instrument may be appropriate include RA, OA, and SSc.

Caveats and cautions. There is no summative total score, which is a disadvantage when using the AHFT in research, as numerous correlations or comparisons must be made for each of the 11 items and not just 1 score. As the AHFT is a performance-based test with many items, numerous items are needed. However, most are easily available and fit into a 24-inch suitcase. Predictive validity and responsiveness to change have not been documented.

Clinical usability. Psychometric evaluation supports interpretation of scores to make decisions for individuals. The administrative burden limits clinical use as the AHFT does need equipment and training and takes time to administer. The respondent burden might limit clinical use as it takes 20–30 minutes for the test.

Research usability. The psychometric evaluation supports research use. However, the administrative burden might limit research use for the same reasons listed under clinical usability. The respondent burden might limit research use, although once one is familiar with the test, it can be administered quickly. As stated above, cutting the putty can be difficult.

AUSTRALIAN CANADIAN OSTEOARTHRITIS HAND INDEX (AUSCAN)

Description

Purpose. The AUSCAN is a self-report measure to assess hand pain, stiffness, and hand function in persons with osteoarthritis (OA) (7).

Content. There are 3 scales: pain, stiffness, and function. Pain is assessed at rest and during activities, including gripping, lifting, turning, and squeezing objects, while stiffness refers to morning stiffness upon waking. The function items ask about difficulty turning, fastening, opening, carrying, grabbing, and squeezing various objects (7).

Number of items. There are 15 items divided into 3 subscales: pain (5 items), stiffness (1 item), and function (9 items).

Response options/scale. Likert scale format from 0 (none) to 4 (extreme); 100-mm visual analog scale (VAS) format from 0 (none) to 100 (extreme).

Recall period for items. The last 48 hours.

Endorsements. None.

Examples of use. See references list.

Practical Application

How to obtain. A copy can be obtained from the web site (http://www.auscan.org) or from Dr. Bellamy (e-mail: n.bellamy@uq.edu.au). It is copyrighted.

Cost. Unknown.

Method of administration. Patient- or clinician-completed questionnaire.

Scoring. There are specific scoring instructions. Computer scoring is not necessary. There are no instructions for handling missing values.

Score interpretation. Score range: on VAS, pain (0–500), stiffness (0–100), and physical function (0–900). Lower scores indicate better status. Normative values are not available.

Respondent burden. Time to complete is ~7 minutes.

Administrative burden. Time to administer is ~7 minutes. Time to score is ~5 minutes. No training is necessary.

Translations/adaptations. Languages available: English, Spanish, French, German, Norwegian, Dutch, and Italian.

Psychometric Information

Method of development. Items were generated from interactions between experts (health providers) and interviews with patients. Items were retained that had a prevalence of >60% in the sample population and a mean importance rating >2.0 (scale from 1–5) (7). Patients were involved in development of the questionnaire. The development of the subscales is not described well. Item-response theory was not used in development or item selection.

Acceptability. Missing data are common. Floor or ceiling effects are possible.

Reliability. *Evidence for internal consistency.* Cronbach's α = 0.90–0.98. Cronbach's alphas for the Likert scale ranged from 0.90–0.94, and for the VAS ranged from 0.94–0.98 (8). For 17 patients with rheumatoid arthritis (RA), Cronbach's alpha for the total score was 0.94 (9). A later study of patients with familial hand OA showed Cronbach's alphas from 0.93–0.96 (10).

Cronbach's alpha was calculated for the total score and subscales (and subgroups sex, race, presence of hand pain, and radiographic hand OA) among a large community sample of 1,730. Cronbach's alphas ranged from 0.89–0.96 (11).

Evidence for stability (test–retest). Test–retest reliability was established by administering the scales 2 times 1 week apart. Intraclass correlation coefficients (ICCs) ranged from 0.70–86 for the Likert scale and from 0.94–0.98 for the VAS (8).

For 17 patients with RA, test–retest reliability was established by administering the scales 5 days apart. ICCs ranged from 0.92–0.93 for the subscales and 0.94 for the total scales (9).

In a large study of 128 patients with OA, weighted kappa coefficients for each item ranged from 0.29–0.77, with an ICC of 0.87 calculated for the total score (12).

Validity. *Evidence of content validity.* Patient-centered development of the items.

Evidence of convergent validity. In OA, correlation coefficients between measures of grip, pinch, pain, global function, physician-rated severity, Functional Index of Hand OA (FIHOA), and the Health Assessment Questionnaire for the Likert scale ranged from 0.33–0.82 and for the VAS ranged from 0.51–0.86 (8).

For patients with RA, physical function subscale scores correlated with the Sequential Occupational Dexterity Assessment (r = 0.81) and the pain scale correlated with the Michigan Hand Outcomes Questionnaire pain scale (r = 0.68) (9).

The AUSCAN function scale with grip and pinch strength (r = 0.63–0.79) and the pain subscale correlated with a single item (pain measure; r = 0.55–0.58) in patients with familial hand OA (10).

For patients with OA, correlations between the FIHOA and AUSCAN pain scales ranged from 0.67–0.74; AUSCAN stiffness scales ranged from 0.44–0.54 and AUSCAN function scales ranged from 0.76–0.86 (8). Another study correlating the AUSCAN and the FIHOA showed correlations of 0.66 for the pain subscale, 0.54 for stiffness, and 0.81 for function (13).

A third study also reported strong correlations between the total AUSCAN and FIHOA (rs = 0.76) and subscales of pain (rs = 0.79), stiffness (rs = 0.58), and physical function (rs = 0.88). This study also showed that the AUSCAN total correlated with hand strength (rs = 0.50) and the Arthritis Impact Measurement Scales 2 physical, arm, and hand (rs = 0.73, 0.63, and 0.69, respectively); the Health Assessment Questionnaire (rs = 0.63); and the Short Form 36 physical component (rs = 0.43) and mental component (rs = 0.08). Correlations with the subscales on the AUSCAN and the measures above were also calculated with lowest correlations with the stiffness subscale (12). Factor analysis supported the pain and function subscales (10).

Ability to detect change. Standardized response means (mean difference between end of the washout for discontinuing the current nonsteroidal antiinflammatory drug and followup at 1, 3, and 6 weeks) for the AUSCAN ranged from −0.74 to −0.23 for the Likert scale and from −0.84 to −0.39 for the VAS (8).

Each 1-unit increase for the function subscale was associated with a clinically relevant decrease in hand strength

(SE ranged from 0.03– 0.11 for the right hand and from 0.04 – 0.11 for the left hand) (10).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The AUSCAN measures hand function, an important aspect of life affected by rheumatic disease. The measure is appropriate for evaluating interventions. Groups or conditions for which the instrument may be appropriate are persons with RA and OA, as most of the psychometrics have been done with these populations.

Caveats and cautions. Because the AUSCAN is copyrighted, it is a little more difficult to obtain than other measures of hand function. Gaps or limitations in psychometric evaluation are with rheumatic conditions other than RA or OA.

Clinical usability. Psychometric evaluation supports interpretation of scores to make decisions for individuals. Neither the administrative nor respondent burden would limit clinical use.

Research usability. The psychometric evaluation supports research use. Neither the administrative nor respondent burden would limit research use.

COCHIN HAND FUNCTION SCALE

Description

Purpose. The purpose of this self-report scale is to measure functional ability in the hand.

Content. The questions ask how much difficulty the person has performing 18 tasks without the help of any assistive device. Kitchen tasks include holding a bowl and a plate full of food, pouring liquid, cutting meat, and peeling fruit. The dressing items include buttoning and opening/closing a zipper. The hygiene items include squeezing a tube of toothpaste and holding a toothbrush. Office items include 2 writing tasks, while other items include turning a doorknob, cutting with scissors, and turning a key in a lock.

Number of items. There is a total of 18 items with 5 subscales: kitchen (8 items), dressing (2 items), hygiene (2 items), office (2 items), and other (4 items).

Response options/scale. 7-point scale from 0 (without difficulty) to 5 (impossible).

Recall period for items. Not applicable (N/A).

Endorsements. N/A.

Examples of use. See references below.

Practical Application

How to obtain. A copy can be obtained from Duruoz MT, Poiraudeau S, Fermanian J, Menkes C, Amor B, Dougados M, et al. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. J Rheumatol 1996;23:1167–72 (14). There is no web site reference and no cost.

Method of administration. Patient- or clinician-completed self-report questionnaire.

Scoring. Scores for each subscale are summed to yield subscale scores and scores from the subscales are summed to yield a total score. Total scores range from 0 –90. Scores for the kitchen subscale range from 0 – 40 and scores for the dressing, hygiene, and office subscales range from 0 –10. Scores for other range from 0 –20. Computer scoring is not necessary. There are no instructions for handling missing values.

Score interpretation. Score range is from 0 –90. A higher score indicates greater disability or more difficulty, whereas a lower score indicates less disability or difficulty. No normative values are available.

Respondent burden. Time to complete is 3–5 minutes. The items are easy to read.

Administrative burden. It takes less than 3–5 minutes to administer and score. No training is necessary.

Translations/adaptations. Languages available are French, English, and Italian.

Psychometric Information

Method of development. Content or face validity was determined by collecting a list of hand activity questions from published indices. The questions were divided into 5 categories and given to 10 subjects. These subjects added other items and evaluated items for clarity. It was then administered to 102 subjects. Questions that were "never done" by >5% of subjects were eliminated, yielding 18 items (14). Patients were involved in development of items. Items were grouped by content to generate subscales. Item-response theory was not used in development or item selection.

Acceptability. Items are easy to read. Missing data are not common. Floor or ceiling effects are possible.

Reliability. *Rheumatoid arthritis (RA).* Intrarater reliability was established by having the same rater interview 25 subjects 2 times 24 hours apart (intraclass correlation coefficient [ICC] 0.97) (14). Interrater reliability was established by having 2 raters interview 68 subjects at 24-hour intervals (ICC 0.96) (14). Test–retest reliability was established by administering the scales 2 times 1 week apart (ICC 0.89) (15).

Osteoarthritis (OA). Interrater reliability was established by administering the scale 2 times within 1 hour to 41 subjects (ICC 0.96) (6). Test–retest reliability was established by administering the scales 2 times 1 week apart (ICC 0.94) (6).

Systemic sclerosis (SSc). Test–retest reliability was established by administering the scales 2 times 1 week apart (ICC 0.97) (16).

Validity. *Convergent validity: RA.* Scores on the Cochin Scale were correlated with scores on a visual analog scale for functional handicap (rs = 0.77) (14). In another sample of subjects with RA, scores were correlated with scores on the Arthritis Hand Function Test (rs = 0.36 – 0.54), Health Assessment Questionnaire (HAQ; rs = 0.78), Scleroderma Functional Assessment Questionnaire (rs = 0.85), and Hand Mobility in Scleroderma Test (rs = 0.39) (15).

Convergent validity: SSc. Scores on the Cochin Scale were correlated with scores on the Arthritis Hand Function Test (r = 0.34 – 0.58), Keital Function Test (rs = 0.48), and HAQ (rs = 0.79) (16).

Construct validity: RA. Scores on the Cochin Scale correlated with scores on the Revel Functional Index (rs = 0.91) and Hand Functional Index (rs = 0.58) (14).

Convergent validity: OA. Scores on the Cochin Scale correlated with scores on the Revel Functional Index (rs = 0.86), Dreiser Functional Index (rs = 0.87), and a visual analog scale to assess perceived disability (rs = 0.67) (17).

Scores on the Cochin Scale also correlated with other self-reports of hand function, including the Michigan Hand Outcomes Questionnaire (rs = 0.82) and Dreiser's Functional Index for Hand OA (rs = 0.89) (8), as well as hand strength (rs = 0.57–0.64) and dexterity (rs = 0.52–0.57) (6).

Divergent validity: RA. Scores on the Cochin Scale were correlated with variables known to have little correlation with disability: age (rs = 0.38), disease duration (rs = 0.23), morning stiffness (rs = 0.41), elbow and shoulder pain (rs = 0.48), hand pain (rs = 0.52), tenderness (rs = 0.51), and swelling (rs = 0.12) (14).

Divergent validity: OA. Scores on the Cochin Scale were correlated with variables known to have moderate or little correlation with disability: Richie articular index (rs = 0.51), visual analog scale for pain (rs = 0.54), clinical impairment (rs = 0.32), and Kallman index score (rs = 0.14) (17).

Construct validity: SSc. Scores on the Cochin Scale explained 75% of the variance on the HAQ (18).

Convergent validity: SSc. Scores on the Cochin correlated with the HAQ (rs = 0.75), scleroderma HAQ (rs = 0.81), Kapandji Index (rs = 0.63), Hand Function Index (rs = 0.58), Short Form 36 (SF-36) physical component score (rs = 0.45), and McMaster Toronto Arthritis Patient Preference Disability Questionnaire (rs = 0.48) (18).

Divergent validity: SSc. Scores on the Cochin Scale had little to no correlation with measures of concepts differing from hand function: anxiety (rs = 0.16), SF-36 mental component score (rs = 0.14), depression (rs = 0.05), disease duration (rs = 0.15), and age (rs = 0.01) (18).

The validity of using the a priori subscales (kitchen, dressing, etc) scores is not supported by exploratory factor analysis in RA, OA, and SSc; 3 factors were identified in RA (14) and OA (17) and 2 factors in SSc (18).

Ability to detect change. *RA.* Fifty-five subjects completed the scale 2 times ~15 months apart. Changes in scores correlated with subject-perceived handicap (rs = 0.58), but had little correlation with disease activity measures (rs = 0.19–0.34) (19).

The responsiveness of the Cochin Scale after surgery was assessed by testing 52 subjects who were going to have wrist and/or finger surgery 48 hours before the surgery and at least 6 months after surgery. Cochin Scale scores significantly improved at the send visit ($P < 0.0001$) with standardized response mean and effect size values of 0.66 and 0.58, respectively (20).

OA. Fifty-one subjects completed the scale 2 times approximately 5 months apart. Changes in scores correlated with subjects' overall assessment (rs = 0.47) (17). The scale also discriminated between those who improved and those who deteriorated ($P < 0.0001$) (17).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument measures hand function, an important element of disease or aspect of life that may be affected by disease. The measure is appropriate for evaluating interventions. Groups or conditions for which the instrument may be appropriate include RA, OA, SSc, and diabetes mellitus.

Caveats and cautions. Items probably need to be updated to reflect common hand activities such as keyboarding, texting, and cell phones. It would be nice if scores were interpreted in terms of severity of hand dysfunction. Because the original article did not give the scale a formal name, the scale as also been termed the Duruoz Hand Index and/or Hand Function Disability Scale, which causes some confusion.

Clinical usability. Psychometric evaluation supports interpretation of scores to make decisions for individuals. Neither the administrative nor respondent burden would limit clinical use.

Research usability. The psychometric evaluation supports research use in clinical trials (21). Neither the administrative nor the respondent burden would limit research use.

FUNCTIONAL INDEX FOR HAND OSTEOARTHRITIS (FIHOA)

Description

Purpose. The purpose of the FIHOA is to measure hand function in persons with hand osteoarthritis (OA). The original was in French in 1995.

Content. Questions ask about using a key, cutting different objects, lifting, buttoning, using tools, writing, and shaking hands.

Number of items. There are 10 items; no subscales.

Response options/scale. Items are rated on a 4-point scale from 0 (possible without difficulty) to 3 (impossible).

Recall period for items. Not applicable.

Endorsements. No.

Examples of use. See reference list.

Practical Application

How to obtain. A copy can be obtained from the 1995 article (22). There is no cost.

Method of administration. Self-report or physician administered.

Scoring. Scores for each item are summed to get a total score. Computer scoring is not necessary. There are no instructions for handling missing values.

Score interpretation. Score range is from 0–30. Low scores indicate better hand function. A minimum score of 4 or 5 was shown to discriminate symptomatic and nonsymptomatic hand OA patients (22). Normative values are not available.

Respondent burden. Time to complete is 3 minutes. Items are acceptable in terms of reading level.

Administrative burden. Time to administer is ~3 minutes. Time to score is 3 minutes. No training is necessary.

Translations/adaptations. Languages available are: French, English, and Dutch. Cultural adaptations have been done for the Dutch version.

Psychometric Information

Method of development. No clear description is given regarding how the items were generated. Patients were not involved in the development. Item-response theory was not used in development or item selection.

Acceptability. Readability is acceptable. Missing data are common. Floor or ceiling effects are possible.

Reliability. *Evidence for internal consistency.* Cronbach's alpha was 0.85 (22). In a recent study of 128 patients with hand OA, Cronbach's alpha was 0.90 (12) and a study on a Dutch version reported a Cronbach's alpha of 0.89 (13).

Evidence for stability (test–retest). The intraclass correlation coefficient (ICC) for the total score was 0.95 when the questionnaire was administered twice 1 hour apart (22). Kappa values for each item ranged from 0.68–0.87 (22).

Test–retest reliability established by administering the FIHOA 2 times 1 week apart yielded an ICC of 0.74 (6) and 2 times with a 5-day interval yielded an ICC of 0.96 for the total score. ICCs for the 10 items ranged from 0.76–0.96 (13). In another study, weighted kappa coefficients for each item ranged from 0.41–0.77, with an ICC of 0.94 calculated for the total score (12).

Intraobserver reliability. Established having investigators interview patients with OA 2 times 1 hour apart. Correlations between the scores assigned by the investigators were 0.95. The mean ± SD difference in scores was 0.17 ± 1.64, with a coefficient of variation of 9.32% (22).

Validity. For persons with OA, correlations between the FIHOA and Australian Canadian Osteoarthritis Hand Index (AUSCAN) pain scales ranged from 0.67–0.74; AUSCAN stiffness scales ranged from 0.44–0.54; and AUSCAN function scales ranged from 0.76–0.86 (8). Another study correlating the FIHOA and AUSCAN showed correlations of 0.66 for the pain subscale, 0.54 for stiffness, and 0.81 for function (13).

A third study also reported strong correlations between the FIHOA and total AUSCAN (rs = 0.76), and subscales of pain (rs = 0.79), stiffness (rs = 0.58), and physical function (rs = 0.88). This study also showed that the FIHOA correlated with hand strength (rs = 0.58) and the Arthritis Impact Measurement Scales 2 physical, arm, and hand (rs = 0.80, 0.71, and 0.69 respectively); Health Assessment Questionnaire (rs = 0.73); and Short Form 36 physical component (rs = 0.67) and mental component (rs = 0.38) (12).

Scores on the FIHOA also correlated with other self-reports of hand function, including the Cochin Hand Function Scale (rs = 0.89) and Michigan Hand Outcomes Questionnaire (rs = 0.86) (6), as well as hand strength (rs = 0.47–0.57) and dexterity (rs = 0.44–0.46) (6).

Ability to detect change. The standardized response mean (SRM) of the FIHOA over 6 months was 0.58, which was lower than the SRM for a pain visual analog scale (SRM 0.87) (23).

The average SRM over a period of 6 weeks was −0.31 and decreased over time (week 1 versus week 6), and the FIHOA was shown to be less responsive than the AUSCAN (8).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument is a quick self-report of hand function often impacted by rheumatic diseases. The measure is appropriate for evaluating interventions but responsiveness may be a concern. All of the psychometrics have been done on OA but the items are relevant to other rheumatic conditions that affect the hand. There is some evidence for diabetes mellitus.

Caveats and cautions. There is weaknesses of the instrument. Discrepancy in symptoms between 2 hands may make it difficult for patients to score items (13). Responsiveness is not high; psychometric evaluation has not been done with other rheumatic diseases such as rheumatoid arthritis and systemic sclerosis that affect hand function.

Clinical usability. Psychometric evaluation supports interpretation of scores to make decisions for individuals. Neither the administrative nor respondent burden would limit clinical use.

Research usability. The psychometric evaluation supports research use. Neither the administrative nor the respondent burden would limit research use.

GRIP ABILITY TEST (GAT)

Description

Purpose. The GAT is a modification of a general test of hand function based on activities of daily living, the Grip Function Test. The GAT is intended to be a simple and rapid test of hand function for persons with rheumatoid arthritis (RA).

Content. Putting a sock over 1 hand, putting a paper clip on an envelope, and pouring water from a jug.

Number of items. 3 items.

Response options/scale. Timed test.

Recall period for items. Not applicable.

Endorsements. None.

Examples of use. See under references.

Practical Application

How to obtain. The GAT administration and equipment are described in the article by Dellhag and Bjelle (24).

Equipment needed. 25 cm of Tubigrip elasticized tubular bandage (7.5 cm wide for women and 10 cm wide for men), metal paper clip (30 × 10 mm), envelope (11.5 cm × 16 cm), 1 liter water jug with handle, cup (2 dl), and stopwatch.

Cost. 10-meter roll of tubigrip: $55.00 available from hand therapy catalogs.

Method of administration. Performance-based test.

Scoring. Each item is timed in seconds and the times are summed to yield a total GAT score.

Score interpretation. Score range is 5–6 seconds to 2–3 minutes. A GAT score of <20 seconds is considered normal. Higher scores mean decreased hand function. Normative values are not available.

Respondent burden. Time to complete is 5–6 seconds to 2–3 minutes. Item difficulty: items are simple hand items.

Administrative burden. Scoring is immediate. The times for the 3 tasks need to be summed, which would take <1 minute. No training is necessary.

Translations/adaptations. None.

Psychometric Information

Method of development. Items for the GAT were chosen from the items on the Grip Function Test that were found to discriminate between patients with RA and controls and that were sensitive to change in a hand training program (24). Patients were involved in development to determine items that discriminated patients with RA from controls. Item-response theory was not used in development or item selection.

Acceptability. Items are easy to read. Missing data are not common. Floor or ceiling effects are possible.

Reliability. Internal consistency: Cronbach's alpha calculated from testing 52 subjects was 0.65 (24). Intraobserver reliability was 0.99 (24). Interobserver reliability for 2 observers rating 20 subjects was 0.95 (24).

Validity. *Content validity.* Items were selected from the Grip Function Test to represent 4 grip types (24).

Convergent validity. Scores from the GAT correlated with scores on the Health Assessment Questionnaire (HAQ; r = 0.53, $P < 0.001$), grip strength (r = 0.29, $P < 0.05$), self-estimated hand function (r = 0.42, $P < 0.01$), pain with nonresisted motion (r = 0.33, $P < 0.05$), pain with resisted motion (r = 0.46, $P < 0.001$), stiffness (r = 0.32, $P < 0.001$), and the Keital Functional Test (r = 0.42, $P < 0.01$) (24).

Construct validity: known groups validity. All items discriminated between persons with RA and controls ($P < 0.001$) (1). Changes in the GAT scores correlated with change in HAQ scores (r = 0.42, $P < 0.01$) (25).

Ability to detect change. Total scores on the GAT ($P < 0.001$) and item scores ($P < 0.01$, 0.01, and 0.05) were sensitive to change after a hand training program (24).

Subjects with low GAT scores displayed normal or increased safety margins in grip force and the load at the point where an object begins to slip out of the fingers compared to healthy controls, whereas subjects who had higher GAT scores exhibited lower safety margins (26).

The GAT did not appear to be sensitive enough to measure differences in patients with osteoarthritis (OA) who had participated in an education program compared to a control group (27).

The standardized response mean (SRM) showed modest sensitivity (SRM 0.6–0.7) in a sample of RA patients who were followed for 1 year while receiving tumor necrosis factor inhibitors (28).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument measures hand function, an important element of disease or an aspect of life that may be affected by disease. The measure seems to be sensitive for evaluating interventions. Groups or conditions for which the instrument may be appropriate include RA and some evidence for systemic sclerosis (29), but use with OA with caution.

Caveats and cautions. Limited psychometrics; most of the psychometric studies and research using this test have been done by one of the developers of the test. A conflicting finding regarding responsiveness is of concern. A group or condition for which there may be problems is OA. Gaps or limitations in psychometric evaluation are described above.

Clinical usability. Psychometric evaluation does not appear strong enough to support interpretation of scores to make decisions for individuals. Neither the administrative nor respondent burden would limit clinical use.

Research usability. The psychometric evaluation support research use for RA. Neither the administrative nor the respondent burden would limit research use.

JEBSEN HAND FUNCTION TEST (JHFT)

Description

Purpose. The purpose of the JHFT is to assess broad aspects of hand function commonly used in activities of daily living using standardized tasks. It was designed for children ages >6 years and adults who have impairments in the hand(s).

The year of publication is 1969. Since then, a commercial version is available. The commercial version has slightly different sizes of equipment than the 1969 publication, which describes a homemade version.

Content. Tasks simulate activities of daily living (see below).

Number of items. There are 7 items (subscales): writing, turning over 3 × 5–inch cards (simulated page turning), picking up small common objects, simulated feeding, stacking checkers, picking up large light cans, and picking up large heavy cans

Response options/scale. Scales for all items are times in seconds.

Recall period for items. Not applicable (N/A).

Endorsements. N/A.

Examples of use. See reference list.

Practical Application

How to obtain. A description on how to construct the test is in the article by Jebsen et al (30). Commercial versions are available from Sammons Preston, Bolingbrook, IL 60440 (online at www.samonspreston.com).

Cost. ~$350.00.

Method of administration. Performance-based test.

Scoring. Each item is timed. Computer scoring is not necessary. There are no instructions for handling missing values.

Score interpretation. Score range is variable depending on disability. The longer the time required to complete the subscales, the more disability a person has. Subscale scores can be compared to the normative tables according to age and sex (30).

Respondent burden. Time to complete is 10–15 minutes, but can be variable depending on the level of disability in the subjects. Younger children, ages 6–7 years, may take up to 20 minutes (31).

Administrative burden. Time to administer: see above. Time to score is minimal as times are recorded after completing each subscale. No training is needed to administer this test; however, the administrator must be familiar with the test and setup for each subscale. Instructions are included in the test kit; however, the original manuscript by Jebsen et al (30) provides clearer instructions.

Translations/adaptations. Languages available: Chinese version.

Psychometric Information

Method of development. Items were chosen to represent a broad aspect of hand function. Patients were not involved in development. Item-response theory was not used in development or item selection.

Acceptability. Oftentimes it is difficult for people to write with their nondominant hand, so data may be missing for that item. Floor or ceiling effects are not possible.

Reliability. *Evidence for stability (test–retest reliability).* In the original study, 26 adult subjects with stable hand disorders were tested at 2 points in time (r = 0.60–0.99) (30). Later, 5 subjects ages >60 years were also tested at 2 points in time (r = 0.84–0.85) (32). The stability of the JHFT over 3 sessions using 20 healthy women showed that subjects performed faster on each successive session; however, only writing and simulated feeding showed a significant difference (33).

To establish test–retest reliability in children, 20 children with stable hand disorders were tested at 2 points in time, 4–10 days apart (r = 0.87–0.99) (31).

Evidence for interrater reliability. Intrarater reliability was established by having 1 rater test 25 patients with rheumatoid arthritis (RA) on successive days (r = 0.82) (34).

Interrater reliability was established by having 2 raters time and score 25 patients with RA (r = 0.91) (25). In another study, interrater reliability was established by having 2 raters simultaneously time and score 5 subjects who were ages >60 years. Intraclass correlation coefficients ranged from 0.82 –1.00 (32).

Validity. *Convergent validity: RA.* Scores on the JHFT correlated significantly with scores on the Arthritis Impact Measurement Scales (AIMS) dexterity items (r = 0.43), the AIMS activities of daily living items (r = 0.47), the AIMS household activity items (r = 0.58), grip strength (r = 0.56), and the Health Assessment Questionnaire (HAQ; r = 0.37) (34). All of the subscales except writing correlated with the HAQ (r = 0.49–0.55) and joint deformity (r = 0.38–0.63) (35). None of the subscales correlated with pain (35).

One hundred twenty-eight subjects were compared stacking wood (standardized) versus plastic (unstandardized) checkers and picking up 1-inch (standardized) versus 1.25-inch (unstandardized) paper clips. Times were significantly faster for the wood checkers than plastic checkers, but not for the paper clips (36).

Discriminative validity. The JHFT was shown to discriminate between subjects with and without different physical disabilities (30,34). However, the mean times were not statistically significantly different between older subjects with OA compared to the norms reported by Jebsen et al (30,37).

Ability to detect change. Effect sizes (ES) and standardized response means (SRMs) were calculated to assess the responsiveness of the JHFT to clinical change at followup compared with baseline for persons with RA (ES 0.47, SRM 0.49) and carpometacarpal joint arthritis (ES 0.67, SRM 0.66) (38).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The test is a widely used and standardized test of hand function. It is easy and quick to administer and can yield subtest scores or an overall score. The measure is appropriate for evaluating interventions. The JHFT has been used to measure hand function in persons with a wide range of diagnoses, ranging from normal aging to arthritis and stroke.

Caveats and cautions. The norms should be revised using the commercially available version of the test. In addition, the hands are tested separately, yet many tasks of daily living are bilateral, i.e., tying a bow, buttoning.

Content validity has been questioned by Mathiowetz (39), who reported that page turning and simulated feeding do not duplicate the actual tasks. More studies assessing the validity and sensitivity of the test are needed.

The JHFT is less responsive to change compared to other questionnaires of hand function (38).

Clinical usability. Psychometric evaluation does not support interpretation of scores to make decisions for individuals. Neither the administrative nor the respondent burden would limit clinical use.

Research usability. The psychometric evaluation supports research use. Neither the administrative nor respondent burden would limit research use.

MICHIGAN HAND OUTCOMES QUESTIONNAIRE (MHQ)

Description

Purpose. The MHQ measures a person's perception of their hands in terms of function, appearance, pain, and satisfaction. The questionnaire was intended for persons with hand and wrist conditions and injuries, including arthritis.

Content. 6 subscales: overall hand function, activities of daily living (ADL), pain, work performance, aesthetics, and patient satisfaction with hand function.

Number of items. There are 37 items and 6 subscales: overall hand function, ADL, pain, work performance, aesthetics, and patient satisfaction with hand function.

Response options/scale. Items are scored on a 5-point Likert scale from 1 (very good/not at all difficult/always/very mild/very satisfied) to 5 (very poor/very difficult/never/severe/very dissatisfied). Raw scores are converted to a scale from 0–100 according to a scoring algorithm.

Ranges for subscales are: hand function (5–25), unilateral ADL (5–25), bilateral ADL (7–35), work (5–25), pain (0–24), aesthetics (4–20), and satisfaction (6–30).

Recall period for items. The past week.

Endorsements. Not applicable.

Examples of use. See references below. In addition, there are numerous studies of the MHQ with conditions other than arthritis that can be found on the web site below.

Practical Application

How to obtain. University of Michigan, Department of Surgery (online at http://sitemaker.umich.edu/mhq).

Cost. None.

Method of administration. Patient- or clinician-completed self-report.

Scoring. There are specific scoring instructions as there is a scoring algorithm in which raw scores are converted to a range from 0–100. Computer scoring is not necessary, but the MHQ can be computer scored using the algorithm and SAS or Microsoft Excel. Instructions for handling missing values are in the article by Chung et al (40) and on the web site.

Score interpretation. Score range: 0–100. Higher scores indicate better performance in all domains except pain. Normative values are not available.

Respondent burden. Time to complete is ~15 minutes.

Administrative burden. Time to administer is 15 minutes. Time to score is 15–20 minutes, but can be computer scored. Training is necessary only to understand the scoring algorithm.

Translations/adaptations. Available in Dutch, Spanish, Chinese, Japanese, Turkish, German, and Korean. Originally developed for populations with hand conditions.

Psychometric Information

Method of development. Items were generated from a Medline search of questionnaires with items related to upper extremity function, and a group of patients were asked what items they considered important for hand function. This generated 100 items. These items were reviewed by patients, hand therapists, and hand surgeons, which generated the 6 subscales (40). Patients were involved in the development. See above for how subscales were generated. Item-response theory was not used in development or item selection.

Acceptability. Readability seems acceptable. Missing data can be common if a person does not do the activity. Floor or ceiling effects are possible.

Reliability. *Evidence for internal consistency: rheumatoid arthritis (RA).* Cronbach's alphas ranged from 0.75–0.94 in a sample of patients with RA (9). In a group of patients with subluxation of the metacarpophalangeal (MCP) joints, Cronbach's alphas ranged from 0.7–0.90 (41).

Evidence for stability (test–retest): RA. The intraclass correlation coefficients (ICCs) for the subscales for 17 subjects with RA who completed the questionnaire twice within 5 days ranged from 0.58–0.97. The ICC for the total score was 0.95 (9).

Spearman's correlations for the subscales for 128 subjects who completed the questionnaire twice 6 months apart ranged from 0.5–0.79, while correlations for the total score ranged from 0.71–0.75 (41).

Evidence for stability (test–retest): osteoarthritis (OA). Test–retest reliability was established by administering the scales 2 times 1 week apart to 40 subjects with OA. ICCs for the subscales ranged from 0.51–0.93, while the ICC for the total scale was 0.85 (6).

Validity. *Evidence of content validity.* Provided above description of how items were generated.

Evidence of convergent validity: RA. Scores on the MHQ pain scale correlated with the Australian Canadian Osteoarthritis Hand Index (AUSCAN) pain scale (r = 0.68), while the MHQ physical function scale correlated with the AUSCAN physical function scale (r = 0.80) (9).

In a sample of 128 subjects with subluxation of the MCP joints, correlations between MHQ subscales scores and the Arthritis Impact Measurement Scales 2 (AIMS2) ranged from 0.20–0.77, with the highest correlations between the function, ADL, work, pain, and overall MHQ scores and the AIMS2 physical function scale and the MHQ pain and AIMS2 symptom scales. Correlations between grip (r = 0.03–0.34) and pinch strength (r = 0.03–0.47) and the Jebsen Hand Function Test (r = 0.22–0.50) were much lower (41).

Convergent validity: OA. Scores on the MHQ also correlated with other self-reports of hand function, including the Cochin Hand Function Scale (rs = 0.82) and Dreiser's Functional Index for Hand OA (rs = 0.86) (6), as well as hand strength (rs = 0.5–0.65) and dexterity (rs = 0.38–0.48) (6).

Ability to detect change. Standardized response means (SRMs) were high for function (SRM 1.42), ADL (SRM 0.89), aesthetics (SRM 1.23), satisfaction (SRM 1.76), overall score (SRM 1.61), pain (SRM 0.63), and work (SRM 0.47) (41).

Using the MHQ satisfaction scale to determine the minimum clinically important difference (MCID), for persons with RA, MCIDs of 3, 11, and 13 were identified for the pain, function, and ADL subscales, respectively (42).

Effect sizes (ES) and SRMs were calculated to assess the responsiveness of the MHQ to clinical change at followup compared with baseline for patients with RA (ES 1.05, SRM 1.07) and carpometacarpal joint arthritis (ES 1.30, SRM 0.93) (38).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument measures hand function, an important element of disease or aspect of life that may be affected by rheumatic disease and adds information on aesthetics and satisfaction with hand function. The measure is appropriate for evaluating interventions. The groups or conditions for which the instrument may be appropriate include hand and wrist conditions and injuries, including arthritis. There are numerous studies on the psychometric properties of the MHQ for persons with conditions other than arthritis that are beyond the scope of this review.

Caveats and cautions. Psychometric evaluation seems to be thorough.

Clinical usability. The psychometric evaluation supports interpretation of scores to make decisions for individuals. The administrative burden could limit clinical use as it takes to 15–20 minutes score the questionnaire using the algorithm. The respondent burden could also limit clinical use as it takes 15–20 minutes to complete the questionnaire.

Research usability. The psychometric evaluation supports research use. Neither the administrative nor respondent burden would limit research use.

AUTHOR CONTRIBUTIONS

Dr. Poole drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Backman C, Mackie H, Harris J. Arthritis hand function test: development of a standardized assessment tool. Occup Ther J Res 1991;11:246–56.
2. Backman C, Mackie H. Reliability and validity of the Arthritis Hand Function test in adults with osteoarthritis. Occup Ther J Res 1997;17: 55–67.
3. Backman C, Mackie H. Arthritis hand function test: inter-rater reliability among self-trained raters. Arthritis Care Res 1995;8:10–5.
4. Poole JL, Gallegos M, O'Linc S. Reliability and validity of the Arthritis Hand Function Test in adults with systemic sclerosis (scleroderma). Arthritis Care Res 2000;13:69–73.
5. Backman C, Cork S, Gibson D, Parsons J. Assessment of hand function: the relationship between pegboard dexterity and applied dexterity. Can J Occup Ther 1992;59:208–13.
6. Poole JL, Lucero SL, Mynatt R. Self-reports and performance based tests of hand function in persons with osteoarthritis. Phys Occup Ther Geriatr 2010;28:249–58.
7. Bellamy N, Campbell J, Haraoui B, Buchbinder R, Hobby K, Roth JH, et al. Dimensionality and clinical importance of pain and disability in hand osteoarthritis: development of the Australian/Canadian (AUSCAN) Osteoarthritis Hand Index. Osteoarthritis Cartilage 2002;10: 855–62.
8. Bellamy N, Campbell J, Haraoui B, Gerecz-Simons E, Buchbinder R, Hobby K, et al. Clinimetric properties of the AUSCAN Osteoarthritis Hand Index: an evaluation of reliability, validity and responsiveness. Osteoarthritis Cartilage 2002;10:863–9.
9. Massy-Westropp N, Krishnan J, Ahern M. Comparing the AUSCAN Osteoarthritis Hand Index, Michigan Hand Outcomes Questionnaire, and Sequential Occupational Dexterity Assessment for patients with rheumatoid arthritis. J Rheumatol 2004;31:1996–2001.
10. Allen KD, Jordan JM, Renner JB, Kraus VB. Validity, factor structure, and clinical relevance of the AUSCAN Osteoarthritis Hand Index. Arthritis Rheum 2006;54:551–6.
11. Allen KD, DeVellis RF, Renner JB, Kraus VB, Jordan JM. Validity and factor structure of the AUSCAN Osteoarthritis Hand Index in a community based sample. Osteoarthritis Cartilage 2007;15:830–6.
12. Moe R, Garratt A, Slatkowsky-Christensen B, Maheu E, Mowinckel P, Kvien T, et al. Concurrent evaluation of data quality, reliability and validity of the Australian/Canadian Osteoarthritis Hand Index and the Functional Index for Hand Osteoarthritis. Rheumatology (Oxford) 2010;49:2327–36.
13. Wittoek R, Cruyssen BV, Maheu E, Verbruggen G. Cross-cultural adaptation of the Dutch version of the Functional Index for Hand Osteoarthritis (FIHOA) and a study on its construct validity. Osteoarthritis Cartilage 2009;17:607–12.
14. Duruoz MT, Poiraudeau S, Fermanian J, Menkes C, Amor B, Dougados M, et al. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. J Rheumatol 1996; 23:1167–72.
15. Poole JL, Cordova KJ, Brower LM. Reliability and validity of a self-report of hand function in persons with rheumatoid arthritis. J Hand Ther 2006;19:12–7.
16. Brower LM, Poole JL. Reliability and validity of the Duruoz Hand Index in persons with systemic sclerosis (scleroderma). Arthritis Rheum 2004;51:805–9.
17. Poiraudeau S, Chevalier X, Conrozier T, Flippo RM, Liote F, Lefevre-Colau MM, et al. Reliability, validity, and sensitivity to change of the Cochin Hand Functional Disability Scale in hand osteoarthritis. Osteoarthritis Cartilage 2001;9:570–7.
18. Rannou F, Poiraudeau S, Berezne A, Baubet T, Le-Guern V, Cabane J, et al. Assessing disability and quality of life in systemic sclerosis: construct validities of the Cochin Hand Function Scale, Health Assessment Questionnaire (HAQ), Systemic Sclerosis HAQ, and Medical Outcomes Study 36-Item Short Form Health Survey. Arthritis Rheum 2007; 57:94–102.
19. Poiraudeau S, Lefevre-Colau MM, Fermanian J, Revel M. The ability of the Cochin Rheumatoid Arthritis Hand Functional Scale to detect change during the course of disease. Arthritis Care Res 2000;13: 296–303.
20. Lefevre-Colau MM, Poiraudeau S, Fermanian J, Etchepare F, Alnot JY, Le Viet D, et al. Responsiveness of the Cochin rheumatoid disability scale after surgery. Rheumatology (Oxford) 2001;40:843–50.
21. Rannou F, Dimet J, Boutron I, Baron G, Fayad F, Mace Y, et al. Splint for base-of-thumb osteoarthritis: a 12-month multicenter randomized controlled study. Ann Int Med 2009;150:661–9.
22. Dreiser R, Maheu E, Guillou GB, Caspard H, Grouin JM. Validation of an algofunctional index for osteoarthritis of the hand. Rev Rhum Engl Ed 1995;62:43S–53S.
23. Dreiser RL, Maheu E, Guillou GB. Sensitivity to change of the functional index for hand osteoarthritis. Osteoarthritis Cartilage 2000;8: S25–8.
24. Dellhag B, Bjelle A. A grip ability test for use in rheumatology practice. J Rheumatol 1995;41:138–63.
25. Dellhag B, Bjelle A. A five-year followup of hand function and activities of daily living in rheumatoid arthritis patients. Arthritis Care Res 1999;12:33–41.
26. Dellhag B, Hosseini N, Bremell T, Ingvarsson PE. Disturbed grip function in women with rheumatoid arthritis. J Rheumatol 2001;28:2624–33.
27. Hansson EE, Jonsson-Lundgren M, Ronnheden A, Sorensson E, Bjarnung A, Dahlberg LE. Effect of an education program for patients with osteoarthritis in primary care: a randomized controlled trial. BMC Musculoskelet Disord 2010;11:244–50.
28. Eberhardt K, Sandqvist G, Geborek P. Hand function tests are important and sensitive tools for assessment of treatment response in patients with rheumatoid arthritis. Scand J Rheumatol 2008;37:109–12.
29. Roberts-Thomson AJ, Roberts-Thomson PJ. Quantitative and qualitative assessment of hand function and deformity in systemic sclerosis [letter]. Rheumatol Int 2007;27:509–10.
30. Jebsen RH, Taylor N, Trieschmann RB, Trotter MJ, Howard LA. An objective and standardized test of hand function. Arch Phys Med Rehabil 1969;50:311–9.
31. Taylor N, Sand PL, Jebsen RH. Evaluation of hand function in children. Arch Phys Med Rehabil 1973;54:129–35.
32. Hackel ME, Wolfe GA, Band SM, Canfield JS. Changes in hand function in the aging adult as determined by the Jebsen Test of Hand Function. Phys Ther 1992;72:373–7.
33. Stern EB. Stability of the Jebsen-Taylor Hand Function Tests across three test sessions. Am J Occup Ther 1992;46:647–9.
34. Vliet Vieland TP, van der Wijk TP, Jolie IM, Zwinderman AH. Determinants of hand function in patients with rheumatoid arthritis. J Rheumatol 1996;23:835–40.
35. Sharma S, Schumacher HR, McLellan AT. Evaluation of the Jebsen Hand Function Test for use in patients with rheumatoid arthritis. Arthritis Care Res 1994;7:16–9.
36. Rider B, Linden C. Comparison of standardized and non-standardized administration of the Jebsen Hand Function Test. J Hand Ther 1988;1: 121–3.
37. Labi ML, Gresham GE, Rathey UK. Hand function in osteoarthritis. Arch Phys Med Rehabil 1982;63:438–40.
38. Sear ED, Chung KC. Validity and responsiveness of the Jebsen-Taylor Hand Function Test. J Hand Surg 2010;35A:30–7.
39. Mathiowetz V. Role of physical performance component evaluations in occupational therapy functional assessment. Am J Occup Ther 1993; 47:225–30.
40. Chung KC, Pillsbury MS, Walers MR, Hayward RA. Reliability and validity testing of the Michigan Hand Outcomes Questionnaire. J Hand Surg 1998;23A:575–87.
41. Waljee JF, Chung KC, Kim HM, Burns PB, Burke FD, Wilgis EF, et al. Validity and responsiveness of the Michigan Hand Questionnaire in patients with rheumatoid arthritis: a multicenter, international study. Arthritis Care Res (Hoboken) 2010;62:1569–77.
42. Shauver MJ, Chung KC. The minimal clinically important difference of the Michigan Hand Questionnaire. J Hand Surg 2009;34A:509–14.

Summary Table for Hand Function Tests*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
JHFT	Items represent hand activities used in daily tasks (writing, simulated page turning, picking up small objects, simulated feeding, stacking checkers, picking up large light and large heavy objects)	Performance-based test	15–20 min	Administrator must be familiar with test and setup for each subscale. Each item is timed	Lower scores indicated faster times. Norms available	Evidence for interrater and test–retest reliability	Evidence for construct validity	Moderate sensitivity to detect change	Assesses variety of hand tasks and is easy and quick to administer	Test needs to be purchased and administrator needs to be familiar with test items and setup. Measures unilateral hand use. Norms are old
GAT	Simple test based on hand activities used in daily tasks (put sock on hand, put paper clip on envelope, pour water)	Performance-based test	2–3 min	2–3 min	Lower scores indicate faster time and better hand function	Evidence from 1 study for intraobserver, interobserver, and internal consistency reliability	Evidence for content and construct validity	Low to moderate sensitivity to detect change	Quick simple test of hand function	Limited psychometrics, all from 1 study. Few items
AHFT	Hand strength and dexterity (grip and pinch strength, dexterity, applied dexterity, applied strength)	Performance-based test	20 min	Administrator must be familiar with set up and administration. 20 min	Lower scores indicate faster times. Manual lists times for normative sample	Evidence for interrater and test–retest reliability	Evidence for construct validity	No evidence	Assesses a variety of hand tasks, including strength. Also assess bilateral hand function	Items for test need to be purchased and some items need to be fabricated. Administrators need to be familiar with test and administration No total scores, only numerous subscale scores
CHFS	Functional ability in the hand (kitchen tasks, dressing, hygiene, office, other)	Self-report	3 min to complete	Hand scored. 18 items summed to obtain a total score (0–90)	High scores indicate more difficulty	Evidence for interrater and test–retest reliability	Evidence for construct validity	Moderate sensitivity	Easy and quick to administer. Assesses a variety of hand function tasks. Strong psychometric support	Items need to be updated
MHQ	Functional ability in the hand (overall hand function, ADL, pain, work performance, aesthetics, patient satisfaction with hand function)	Self-report	15 min to complete	Hand or computer scored. If hand scored, need to understand the scoring algorithm	Higher scores indicate better performance for all subscales except pain	Evidence for internal consistency and test–retest reliability	Evidence for construct validity	Moderate to high sensitivity	Strong psychometric support, especially for RA and other hand injuries and conditions	Limited use with rheumatic conditions other than RA. Takes longer to complete and score than other self-reports
AUSCAN	Assesses hand function (pain, stiffness, and hand function)	Self-report: Likert scale or VAS	7 min	7 min	Lower scores indicate better function	Evidence for internal consistency and test–retest reliability	Evidence for construct validity and factor analysis supports pain and function subscales	Low to moderate sensitivity	Psychometrics are fairly strong. Two response scales available. Translated into many languages	Copyright harder to obtain
FIHOA	Assesses hand function	Self-report or can be completed as interview	3 min	3 min	Lower scores indicate better hand function	Evidence for internal consistency and test–retest reliability and intraobserver reliability	Evidence for construct validity	Low to moderate sensitivity	Quick self-report of hand function	Responsiveness not as high as other tests. Psychometrics on OA; limited for other rheumatic diseases

* ES = effect size; SRM = standardized response mean; MCID = minimum clinically important difference; JHFT = Jebsen Hand Function Test; GAT = Grip Ability Test; AHFT = Arthritis Hand Function Test; CHFS = Cochin Hand Function Scale; MHQ = Michigan Hand Outcomes Questionnaire; ADL = activities of daily living; RA = rheumatoid arthritis; AUSCAN = Australian Canadian Osteoarthritis Hand Index; VAS = visual analog scale; FIHOA = Functional Index for Hand Osteoarthritis; OA = osteoarthritis.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S98–S111
DOI 10.1002/acr.20598
© 2011, American College of Rheumatology

Measures of Systemic Sclerosis (Scleroderma)

Health Assessment Questionnaire (HAQ) and Scleroderma HAQ (SHAQ), Physician- and Patient-Rated Global Assessments, Symptom Burden Index (SBI), University of California, Los Angeles, Scleroderma Clinical Trials Consortium Gastrointestinal Scale (UCLA SCTC GIT) 2.0, Baseline Dyspnea Index (BDI) and Transition Dyspnea Index (TDI) (Mahler's Index), Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR), and Raynaud's Condition Score (RCS)

JANET POPE

INTRODUCTION

Outcome measurements are important in evaluating patients with systemic sclerosis (SSc; scleroderma) and in research such as clinical trials, and many patient-reported outcomes can be useful for monitoring SSc patients seen in practice.

Previous research and consensus exercises have demonstrated important domains that may be useful in SSc clinical trials. These include skin, musculoskeletal, cardiac, pulmonary, cardiopulmonary, gastrointestinal, renal, Raynaud's phenomenon and digital ulcers, health-related quality of life and function, global health, and biomarkers (1,2). This review will focus on the Health Assessment Questionnaire (HAQ) disability index in SSc (3) and the Scleroderma HAQ (SHAQ) (4), physician and patient global assessments, the University of California, Los Angeles, Scleroderma Clinical Trials Consortium Gastrointestinal Scale (UCLA SCTC GIT) instrument for gastrointestinal involvement in SSc (5), the Raynaud's Condition Score (RCS) (6), Symptom Burden Index (7), the Cambridge Pulmonary Hypertension Outcome Review scale for pulmonary arterial hypertension (8), and some dyspnea scales for SSc lung disease, which have partially been validated in SSc.

Fatigue scales, general quality of life measurements, and several hand scales (Cochin Hand Function Scale, the Duruoz Hand Index, the Disabilities of the Arm, Shoulder, and Hand Questionnaire, Arthritis Hand Function Index, Italian Hand Mobility Scale, and the Delta Finger-to-Palm measure) will be discussed elsewhere in this supplement and are not part of the SSc review. In addition, outcomes used in the assessment of SSc for research and/or clinical care such as skin scores, pulmonary function tests, echocardiogram, functional class, 6-minute walk distance, renal measurements, digital ulcer burden, pulmonary imaging, inflammatory markers, joint counts, time to clinical worsening, and disease activity and damage scales are excluded from this review. Measurement of depression and comorbidities are value added in certain circumstances within SSc. All these instruments may be important to consider in a complete review of SSc measurement scales. The durometer and digital ulcer outcomes were not reviewed.

A literature search was performed on July 1, 2011 using PubMed for key words including validity, reliability, and questionnaire, and combining with HAQ, SHAQ, global assessments, SSc or scleroderma, GI outcomes, UCLA GIT, quality of life (QOL), Raynaud's Condition Score (RCS), and dyspnea scales for determining the characteristics of the outcome measurement tools that were within the scope of the study. The abstracts of original and review articles were read, and those thought to be relevant to this topic were fully read to extract instrument characteristics such as reliability, validity, and minimal important difference or minimum clinically important difference. Scales that are reported elsewhere in this supplement, such as general quality of life and other versions of the HAQ, were not reviewed, except for some modifications for SSc. Many clinical measures that are important in SSc, such as skin score and disease and damage indices, were not reviewed.

The results of the search included the following: n = 41 for scleroderma and questionnaire and validity, n = 29 for scleroderma and Health Assessment Questionnaire and validity, n = 14 for HAQ and scleroderma and reliability, n = 9 for HAQ and SSc and reliability, n = 0 for SHAQ and SSc and reliability, n = 1 for SHAQ and scleroderma and reliability, n = 14 for Scleroderma HAQ and scleroderma and reliability, n = 3 for UCLA GIT and SSc, n = 61 for

Janet Pope, MD, MPH, FRCPC: University of Western Ontario, London, Ontario, Canada.

Address correspondence to Janet Pope, MD, MPH, FRCPC, 268 Grosvenor Street, London, Ontario, N6A 4V2, Canada. E-mail: janet.pope@sjhc.london.on.ca.

Submitted for publication March 2, 2011; accepted in revised form August 2, 2011.

pain assessment and SSc, n = 100 for pain assessment and scleroderma, n = 52 for global assessment and scleroderma, n = 38 for global assessment and SSc, n = 12 for Raynaud condition score, and n = 19 for Raynaud's condition score. Symptom burden index and scleroderma resulted in 3 articles. Articles that discussed validation of the selected instruments were included if they were within the scope of this review and not reported elsewhere in this supplement.

MEASURES OF FUNCTION IN SYSTEMIC SCLEROSIS (SSc; SCLERODERMA)

HEALTH ASSESSMENT QUESTIONNAIRE (HAQ) AND SCLERODERMA HAQ (SHAQ)

This section will concentrate on the HAQ disability index (HAQ DI) within the context of SSc and SSc modifications.

Description

The HAQ DI is a self-reported questionnaire in 8 domains. The SHAQ consists of the HAQ (8 domains) and also includes the following scales: pain, patient global assessment, vascular, digital ulcers, lung involvement, and gastrointestinal involvement (4).

Purpose. The HAQ measures self-reported function and is one of the most commonly used quality of life measures in SSc. Due to the multisystem nature of SSc, it can greatly impact a patient's functioning and quality of life. Patient-centered outcomes are important in both clinical practice and research studies.

Developers. Fries et al developed the HAQ, which has been used extensively in SSc (3). Steen and Medsger added the visual analog scales (VAS) to the HAQ to create the SHAQ (4). An excellent review on the measurement properties of the HAQ and SHAQ has been published (9). The HAQ and SHAQ have been extensively studied for clinimetric properties in SSc (4,6,9,10–33).

Scoring. *HAQ.* Scoring is fast and each question is scored 0–3 (where 0 = without difficulty and 3 = unable to do) (3). There are 8 categories and the maximum from each category is added together and divided by the number of categories completed. There is an added point, to a maximum of 3, in each category if aids/devices are checked as being used (if the score is already at 3 or "unable to do," then the score cannot increase further).

SHAQ. The SHAQ is scored like the HAQ, and the other domains are continuous VAS instruments that are measured and then changed to a 0–3 scale. Each area is scored, and the scores are not added together for the VAS components. Therefore, a score can be 1.25 on the HAQ, 0.5 on pain, and separate scores for each of the other items.

Reliability. A therapist observed and graded the activities of the HAQ in patients with SSc. The therapist versus patient intraclass correlation coefficients (ICCs) ranged from 0.38–0.76, and there were significant differences between the observer and subjects' responses for 4 items, whereas the other items had moderate to good agreement (13). Since the HAQ and SHAQ are self-reported, the agreement between observed versus the patient does not need to be high since the instrument is not meant to be scored by an observer. Merkel et al demonstrated reliability of the HAQ and other VAS measures including Raynaud's phenomenon–specific scales in an analysis of data from a large Raynaud's phenomenon trial in SSc (6). Very good within-patient test–retest reliability, if stable SSc patients completed the HAQ once and then 2 days later, has been documented (33). Therefore, overall the HAQ in SSc seems reliable.

Validity. Cole et al compared the HAQ in SSc and early rheumatoid arthritis (ERA), and structural validity was demonstrated (25). HAQ scores have also been compared between established RA and SSc, where the HAQ is on average higher in SSc (21). Convergent and construct validity have been shown with a strong correlation (r = 0.9) between the HAQ and the UK Scleroderma Functional Score (UKFS) (17). There is face and content validity. The mean HAQ is higher in diffuse cutaneous (dcSSc) than limited cutaneous (lcSSc) (32). HAQ is correlated with skin scores, joint pain, tendon rubs, contractures, grip strength, thumb abduction, wrist extension, and motion of the index and middle fingers, and in some studies, the presence versus absence of digital ulcers, but not another (4,6,10,32). Higher HAQ scores are related to more work disability (21). HAQ and pain were found to be related to the physical component score of the Medical Outcomes Study Short Form 36 (SF-36; R = 0.70) in a cross-sectional study of 89 patients with SSc (27).

Predictive validity. A low baseline HAQ score is predictive of improving skin scores over the next year in 2 early dcSSc trials. There was a 1.5 to 5-fold chance of improvement in skin scores and patient global assessments if there was a low baseline HAQ at trial entry in early dcSSc (18). A low HAQ score is also predictive of improved patient global assessment the following year in SSc patients followed in a clinic setting (31). Another study found the HAQ to be correlated with skin, cardiac and renal involvement, tendon friction rubs, hand contractures, proximal muscle strength, and survival in SSc longitudinally (4).

Two large cohorts (one with ERA and the other established SSc) were compared, and there was structural validity in comparing the HAQ scores between the 2 groups (25). SSc patients with joint involvement had higher HAQ scores than in psoriatic arthritis, whereas pain was higher in SSc than RA (26).

Ability to detect change. In a 6-month randomized controlled trial of dcSSc, there was good agreement with the HAQ if skin score improved by at least 30%, with ICCs ranging from 0.69–0.91 (good to excellent). The SF-36 had a larger magnitude of responsiveness for physician and patient global assessments compared to the HAQ, whereas the HAQ was more responsive for skin score and the forced vital capacity on pulmonary function testing (19).

A change in digital ulcers status was related to a change in the HAQ and this was statistically significant when 2 digital ulcer trial results were combined in an exploratory analysis (29). A nonvalidated HAQ subscale score, which contained items primarily asking about finger function, demonstrated an improvement with treatment used to prevent digital ulcers, but this was a post hoc analysis and using a part of the validated scale may not be appropriate, so this subscale needs further validation (28).

The minimum important difference (MID) of the HAQ has been calculated in clinical practice combining patients with lcSSc and dcSSc. MID estimates for improvement and worsening, respectively, were −0.0125 (for the mean, which is well below any measurement that is detectable, or a change of −0.125 if the 75th percentile was used, which has a more reasonable estimate) and 0.042 (for the mean change in the worsened group or worsening by 0.217 if the 75th percentile was used) within one SSc clinic between followup visits. In the Canadian Scleroderma Research Group, where patients have data collected annually, MID estimates for improvement and worsening were −0.037 (−0.250, 75th percentile) and 0.140 (0.375, 75th percentile), respectively (22). This method of MID calculation gives a HAQ that is well below a minimal change in the instrument (which is 0.125, so the 75th percentile was studied). This could be the case due to patients getting worse or better in areas that are not related to function, such as dyspnea or GI symptoms, and this would not be expected to affect the HAQ. Whereas, the MIDs in early dcSSc from the D-penicillamine trial were −0.10 for HAQ improvement and 0.14 for worsening (0.15–0.21 effect size) (23). The MID determined by a Delphi of SSc experts was estimated to be 0.2–0.25 for the HAQ (34). Therefore, depending on how the MID is calculated (or in the latter case estimated by experts to be relevant), the results may be different. It is likely that the MID of the HAQ in SSc, when function is changing, is in reality between 0.125 and 0.25 (1- or 2-point differences on the HAQ scale). Disability as measured by the HAQ worsens over time in SSc by 0.039 (95% confidence interval [95% CI] 0.018–0.061) to 0.071 (95% CI 0.048–0.094) per year, or at least on average 0.12 over 3 years (24).

SHAQ validity and reliability. When compared to the UKFS, the SHAQ had concurrent and convergent construct validity (17). It has face validity (4,6). Reliability of the HAQ and SHAQ VAS was demonstrated using data from a Raynaud's phenomenon trial (6).

Alternate scoring of SHAQ. In one study, the HAQ appeared more reliable than the SHAQ if the scales were added, but this is not routinely done. Using the French translation of the SHAQ, Georges et al proposed a combined score, obtained by pooling the 8 domains of the HAQ DI and the 5 VAS scales, and called it the SSc HAQ. However, this approach has not yet been widely accepted (35).

SHAQ predictive validity. The VAS subscales of the SHAQ were significantly correlated with objective parameters (4). Regarding convergent and construct validity, the SHAQ should have further face and content validity over the HAQ since it includes SSc-specific manifestations (9).

SHAQ responsiveness to change. The SHAQ was responsive to change in a cohort and in a Raynaud's phenomenon trial in SSc (4,6).

HAQ compared to SHAQ. The HAQ was compared to VAS scales of the SHAQ and the UKFS (17). The HAQ and UKFS were strongly correlated (r = 0.9) and both tools were significantly related to other clinical measures. Not surprisingly, the correlations with the VAS were not as strong since they were compared with a functional scale, which would not be as relevant to VAS scales of various organs or symptoms (17).

The SHAQ was no better than the HAQ in discriminating between lcSSc and dcSSc (20). The lung VAS has incremental concurrent validity over the HAQ as an outcome measure evaluating SSc lung disease (36).

Validation of HAQ for SSc and SHAQ in other languages. The SHAQ has been validated in French-speaking SSc patients for structural and convergent validity, with strong coefficients between the HAQ and the physical component score of the SF-36 (r = −0.74, $P < 0.0001$). Discriminant validity was found as the HAQ separated dcSSc and lcSSc (worse in the former). The test–retest reliability was excellent (r = 0.98) (36). The HAQ has been translated into Japanese. In Japanese SSc patients, the HAQ was related to many other clinical variables, especially hand extension. The mean HAQ was lower than what has been reported in US patients with SSc (37). The HAQ has face validity in Japanese (38). The HAQ has been translated into Italian, and significant differences in the HAQ were found in those with higher versus lower modified Rodnan skin thickness scores (above and below 14 units, mean ± SD HAQ in former of 1.158 ± 0.176 versus 0.652 ± 0.076; $P < 0.001$). When present, other clinical features (contractures, myopathy, and digital ulcers) had higher HAQ scores than if absent (39).

Critical Appraisal of Overall Value to the Rheumatology Community

Scoring the HAQ includes adding aids and devices. However, in SSc a change from a low HAQ denoting little or no disability to moderate disability occurs when aids/devices are scored in the total HAQ score in the latter but not the former scenario. This must be taken into account when describing cohorts or trials and the method of scoring (with or without aids and devices should be stated) (40). However, the usual scoring is adding the aids and devices as mentioned above.

The HAQ is widely used, inexpensive, and takes only a couple of minutes to complete and score. It has been translated into many languages, with some validation in SSc in other languages. The HAQ is somewhat outdated and may not apply to patients in different countries (such as opening a milk carton, lifting a certain amount in pounds, taking a tub bath, etc.).

OTHER MEASURES OF FUNCTION IN SSc

There are strong correlations between the UK Scleroderma Functional Score (UKFS) and the Health Assessment Questionnaire (HAQ) in a cross-sectional SSc sample ($\rho = 0.90$, $P < 0.0001$) and prospectively with change over time comparing UKFS and HAQ ($\rho = 0.59$, $P < 0.0001$) (41). The Functional Index, which is an 11-item scale, has not been widely used in SSc (42).

The Scleroderma Assessment Questionnaire is a self-assessed measure ranging from 0–3 for several questions including vascular, respiratory, gastrointestinal, musculoskeletal, and overall disease status with 23 questions divided into 4 groups. The questions include the HAQ and other questions (43,44). It has face validity and is sensitive to change, but currently is not commonly used. Another measure that has not been validated is the Systemic Sclerosis Questionnaire, which includes general, organ-specific, and musculoskeletal complaints (45). There are other proposed functional scales (46–48).

GLOBAL ASSESSMENTS IN SYSTEMIC SCLEROSIS (SSc; SCLERODERMA)

PHYSICIAN- AND PATIENT-RATED GLOBAL ASSESSMENTS

Description

Global assessments are rated by the observer (usually a physician) or the patient. They may be rated from 0–100 with a continuous visual analog scale (VAS) or a Likert scale with, for example, a 10-point rating or a change scale, such as a 7-point scale (from −3 to +3 including 0 in the center for no change). There is no standardization for the scale, but usually a low number indicates less disease activity. They can also rate severity, damage, or overall disease. Neither the question or the recall period is standardized (e.g., a global assessment may ask the patient to rate overall disease activity or the way that SSc has affected her/him over the last month, week, or today since the last visit).

Purpose. The global assessments are very easy and are used in both clinical practice and research studies to quantify the disease activity or severity (or whatever is being asked). The most frequent rating is SSc overall disease activity.

Content. There may be one scale (e.g., an overall global assessment) or several scales (e.g., organ areas or Raynaud's phenomenon symptoms).

Developer. None.

Number of items in scale. There may be 1 global assessment each for the patient and physician to complete, and/or an additional series of global assessments of each organ system. There may be questions with respect to activity, damage, and severity.

Scoring. Scoring is easy and may be a 10-cm or 15-cm VAS and converted to 0–100 for the former or 0–3 for the latter. Likert scales do not usually have measurement in between the numbers provided such as 0,1,2–9,10. It takes only a few seconds to complete and to score.

Reliability. Test–retest reliability has been calculated for physician and patient global assessments (49). Measures with inherent interpretation such as global assessments and skin scores were found to have increased variability than easily-performed measurements such as grip strength and oral opening. However, there was good reproducibility within observers, but moderate between observers' intraclass correlation coefficients (ICCs). The patient global assessment has very good reproducibility (33). The relatedness of physician global assessments in SSc for disease activity, severity, and damage has been calculated ranging from $\rho = 0.77$ for severity and activity to $\rho = 0.61$ for damage and activity (50).

Validity. There is some face validity since a patient or rater is asked to determine overall disease activity or severity. The question may be open to interpretation. The ratings are different between the patient and the physician, so they are measuring different things. There are higher scores in general for diffuse cutaneous SSc (dcSSc) versus limited cutaneous SSc (lcSSc) (50). A large Canadian database (Canadian Scleroderma Research Group) and a Michigan database demonstrated that there is discordance between the patient and physician global assessments. Patients perceived greater disease severity than physicians (mean ± SD difference 0.78 ± 2.65). The agreement between patient and physician assessments of disease severity was modest (ICC 0.38, weighted $\kappa = 0.38$). Both patient and physician scales were related to skin scores, breathlessness, and pain, but the relative importance of these predictors differed. Patients were also influenced by other subjective symptoms, while physicians were also influenced by disease duration and creatinine. The predictors explained 56% of the deviance in the patient global assessments and 29% in the physician assessments (51). This makes sense as they are not measuring the same things and are both necessary end points for measuring the status of a patient with SSc. Disease activity was rated higher for dcSSc (especially early dcSSc) compared to lcSSc (50).

Sensitivity to change. Low Health Assessment Questionnaire (HAQ) scores predicted improvement in the physician global assessment in clinical practice over 1 year (31). In the D-penicillamine trial, multivariate logistic regression demonstrated that the physician's global assessment of improvement was best explained by a model with skin score and HAQ ($R^2 = 0.46$) (52). Skin scores and patient global assessments were correlated with improvement in 2 early dcSSc trials (r = between 0.25 and 0.35) (18). The minimal important differences for patient global assessment in SSc have been calculated in clinical practice and are very small (4–6.7 on 100-mm VAS) (22). Minimum clinically relevant important differences from a physician's perspective, obtained by expert opinion and Delphi consensus, were 3–7.5 units of the modified Rodnan skin thickness score, 8–13 for physician global (out of 100), and, similarly, 10–12 for patient global assessment (out of 100) (34).

Critical Appraisal of Overall Value to the Rheumatology Community

Global measurements are important and easy to use. They have some validity and give a rating to what the patient perceives as important for a patient-reported global assessment. It is difficult for a patient living with a complex disease such as SSc to know the difference between disease activity and damage, so the physician and patient global assessments may be very different (51). The global assessment question depends on what is asked, since standardization is lacking and different questions may have variable sensitivity to change, i.e., severity may or may not change, damage will potentially worsen over time but not improve, and disease activity within a study can change for the better or worse or remain stable. In addition, patients and physicians are not measuring the same things or they are weighing them differently and therefore differences between patient and physician global ratings do occur (51). The physician assessor may score serious organ involvement that is active as worse than mild complications, whereas a patient rates what they are feeling and may score a digital ulcer or gastrointestinal problem higher than a serious organ involvement, especially if the latter has minimal symptoms. Continuous and Likert scales are not completely interchangeable.

PAIN ASSESSMENTS IN SSc

Description

Pain in SSc is most often assessed by visual analog scales (VAS), Likert, or change scales. Questions are not different from other generic pain scales except there could be attribution, such as overall pain from SSc, Raynaud's phenomenon (RP) pain, or digital ulcer pain, or there is no attribution (such as overall pain compared to attribution asking about pain from SSc or from a specific problem such as RP, digital ulcers, skin involvement, arthritis, or gastrointestinal [GI] problems).

Purpose. Self-administered scale to rate SSc-related pain.

Content. Scale or scales on pain.

Developer. None.

Number of items in scale. Single item, part of other scales (such as overall disease, overall pain, or digital ulcer pain), or part of a multiquestion pain questionnaire. The Raynaud's Condition Score (RCS) and Symptom Burden Index contain pain questions. There may be subscales of various pain areas (from RP, ulcers, GI, overall, etc.).

Scoring. The Health Assessment Questionnaire (HAQ) pain scale is 15 cm with a conversion from 0 (no pain) to 3 (100% pain). The measured number on the scale is divided by 5 cm to make it be scaled appropriately for the HAQ pain scale. Other scales could be numbers, Likert, or 100-mm VAS. Completion should be less than a minute. Scoring is also fast.

Reliability. Pain is correlated with other patient-reported outcomes of disease activity in RP with SSc (6). Pain is very common in SSc with mean pain of ~40 out of 100 in a clinic setting (22). In a large SSc study, 83% of patients had pain, half of whom had mild pain (1–4 out of 10); one quarter had moderate pain (5–7), and 10 had severe pain (8–10). More frequent RP attacks, active ulcers, worse synovitis, and GI symptoms were associated with pain. Overall pain was worse in diffuse cutaneous SSc (53). The modified Rodnan skin thickness score is strongly associated with pain (54). Higher pain scores are also associated with more alteration of body image in SSc (55).

Validity. The pain scale is validated in SSc (content) alone and with the HAQ or RCS (4,6). The mean ± SD pain in SSc patients is 41 ± 26 out of 100 (22).

Sensitivity to change. In many effective therapies for RP in SSc, the pain scale improves (56). The minimum important difference (MID) for pain in SSc clinical practice on a 100-mm VAS is from 3.6–8 for pain (22). Physicians perceived the MID for pain in SSc to be 0.2–0.3 (out of 3) (34).

Critical Appraisal of Overall Value to the Rheumatology Community

Pain is likely under-recognized in SSc and is important to measure. There is a lack of standardization for the time frame and actual question in SSc with respect to pain. There can be overall pain as well as organ-specific pain questions, and scales can be 100-mm, 15-cm, Likert, and even descriptive. The same limitations of global assessments apply to the assessment of pain. Pain and attribution from disease under study or other problems are very difficult for patients, and some pain questions may be about disease-related pain while another scale may be about overall pain. Therefore, mechanical back pain would not be included in the former, but it would in the latter. However, the test–retest reliability should not be affected in either scenario, but the attribution to SSc is not necessarily present in a question that asks about overall pain. In addition, even if asked about SSc-specific pain, many patients rank all their pain as they do from their disease. It has been found that patients have problems distinguishing SSc from other comorbidities (57). Pain can be from disease activity or damage and therefore may not be responsive to treatment.

FATIGUE

Fatigue is a very common complaint in SSc, but there are no specific SSc fatigue scales. As with other rheumatic diseases, pain and fatigue are significant determinants of quality of life in SSc (58). The minimal important difference for fatigue in SSc clinical practice is from 3.8–10.0 out of 100, and a sleep problems visual analog scale was from 5.9–18.5 (22). A detailed review of fatigue scales has recently been completed (59).

SYSTEMIC SCLEROSIS (SSc; SCLERODERMA)–SPECIFIC MEASURES OF QUALITY OF LIFE

SYMPTOM BURDEN INDEX (SBI)

This is a self-reported questionnaire for SSc.

Description

The SBI was developed to determine the effects of SSc in several domains that impact quality of life (QOL) (7). The SBI has 8 major symptomatic areas (skin, hand mobility, calcinosis, shortness of breath, eating, bowel, sleep, and pain) (7).

Purpose. The SBI determines the effects of SSc from a patient's perspective in several domains beyond physical function and generic health-related QOL instruments. It is a patient-reported instrument, measuring burden of illness in SSc (7).

Content. The domains consist of several areas with 5 questions in each domain with Likert scales for each question.

Developer. M. A. Kallen and Maria E. Suarez-Almazor, University of Texas.

Number of items in scale. Eight major symptomatic areas of importance to patients are included (skin, hand mobility, calcinosis, shortness of breath, eating, bowel, sleep, and pain), with 5 items each per area with a 0–10 Likert scale. The questions are based on how much, how often, how much interference, how often interfering, and how important is this to the patient.

Scoring. There are 40 questions (5 questions in each of 8 domains). Scoring is done for each scale with the average burden score reported per problem area (in 8 domains) on an 11-point scale (from 0–10). The SBI also gives the number of patients experiencing each SSc-related problem in a group of SSc patients and the number of problems experienced by each patient.

Reliability. Inter-item and item-total score correlations per item set were all moderate to high, and internal consistency reliability estimates were high. These scale characteristics reflect the small to moderate item score ranges observed per item set from 0.4–2.2 (7). Patients had a mean of 5.7 problems with one-third having 0–5 problems and another one-third having 7 or 8 problems in the total of 8 domains.

Validity. The SBI is partially validated in a single site study with 62 SSc patients. Scores in each domain correlated with the Health Assessment Questionnaire (HAQ) and Medical Outcomes Study Short Form 36 (SF-36). For the HAQ, correlations of each SBI scale ranged from 0.3–0.6 and were statistically significant. For the SF-36, higher SBI scores negatively correlated with the SF-36 for both the mental and physical components (7).

Construct validity. Focus groups were tested in order to develop the domains of importance to patients with SSc (60). Except for a few correlations comparing shortness of breath to other domains, all other domains were statistically significantly related to the other domains with low to moderate correlations. However, the burden scores across problems were relatively independent.

Sensitivity to change. Sensitivity to change has not yet been demonstrated.

Translations/adaptations. This has not yet been done.

Critical Appraisal of Overall Value to the Rheumatology Community

The 8 problem areas each have a score and are somewhat independent from the other problems, and SBI scores correlate with the HAQ and SF-36. Especially pain (localized or generalized), fatigue, and malaise were reported to have a major influence on QOL. Gastrointestinal (GI) symptoms were prevalent and had high scores. This could potentially be used in clinical practice and in research. The SBI should be further studied in other cohorts and sensitivity to change is important to determine if the SBI will be used in the future as an outcome measurement in treatment trials. The Scleroderma HAQ visual analog scales for GI, lung, and pain have not been compared with the SBI, where one would expect very strong correlations with the respective scales.

GASTROINTESTINAL (GI) SCALES

The GI tract is a source of considerable discomfort, morbidity, and mortality in patients with SSc. The approach to GI tract–related outcome measures logically follows the pathogenesis, including dysmotility (dysphagia, early satiety, bloating, small bowel bacterial overgrowth, and malabsorption), patent lower esophageal sphincter with gastroesophageal reflux disease, watermelon stomach causing anemia, and obstipation or constipation from large bowel dysmotility, etc. Measures have been validated for manometry and for esophageal and gastric transit time, but these measures may not change significantly in the timeframe of SSc trials. Often, measures are used that have been successful in other GI diseases.

UNIVERSITY OF CALIFORNIA, LOS ANGELES, SCLERODERMA CLINICAL TRIALS CONSORTIUM GASTROINTESTINAL SCALE (UCLA SCTC GIT) 2.0

Description

Khanna et al have validated and improved upon the SSc GIT 1.0, shortening it to the UCLA SCTC GIT 2.0 instrument, which can potentially be used as an outcome for randomized controlled trials in SSc-associated gastrointestinal (GI) involvement (5,61). This is a 7–multi-item scale with areas of reflux, distention/bloating, diarrhea, fecal soilage, constipation, emotional well-being, and social functioning and has been shown to have a good test–retest reliability (5).

Purpose. To have a self-reported GI quality of life (QOL) tool specifically for the range of problems that can occur in SSc and to be able to score the instrument, looking for changes over time or within a trial.

Developer. Dinesh Khanna, et al. University of California, Los Angeles.

Number of items in scale. There are 34 items in the UCLA SCTC GIT 2.0 instrument. The 7 multi-item scales include reflux, distention/bloating, diarrhea, fecal soilage, constipation, emotional well-being, and social functioning.

Scoring. Version 2.0 consists of 34 items scored from 0–3, with lower values indicating better health-related (HR) QOL. The total UCLA SCTC GIT 2.0 score averages 6 of 7 scales (excluding constipation) and is scored from 0 (no GI problems) to 3 (most severe).

Reliability. Test–retest reliability estimates were ≥0.68 (5).

Validity. Self-rated severity of GI involvement has spanned no symptoms to very mild (39%), mild (21%), moderate (31%), and severe/very severe (9%) (5). It is also related to poor sleep (62).

Discriminant validity. Symptom scales were also able to discriminate subjects with corresponding clinical GI diagnoses. The total UCLA SCTC GIT 2.0 score, developed by averaging 6 of 7 scales (excluding constipation), was reliable and provided greater discrimination between mild, moderate, and severe self-rated GI involvement than individual scales.

The 2.0 version was developed using the 52 items from the SSC-GIT 1.0 and 1 rectal incontinence item, grouped into 8 scales based on content: reflux, distention/bloating, diarrhea, fecal soilage (to assess rectal incontinence), constipation, pain, emotional well-being, and social functioning (5). Version 2.0 contains 34 items and is scored from 0–3, with lower values indicating better HRQOL. Therefore, in version 2.0, 7 multi-item scales (reflux, distention/bloating, diarrhea, fecal soilage, constipation, emotional well-being, and social functioning) are included. The UCLA SCTC GIT 2.0 instrument was correlated with depression (except for the parameter of fecal soilage) (63).

Critical Appraisal of Overall Value to the Rheumatology Community

The UCLA SCTC GIT 2.0 scale has been partially validated. It is unknown if this will be sensitive to change in a GI treatment trial. In patients with very frequent symptoms, a moderate improvement may not be detected since the symptoms could still be frequent even if occurring far less often. No comparison of the UCLA SCTC GIT 2.0 questionnaire and the GI visual analog scale on the Scleroderma Health Assessment Questionnaire was found in the literature search.

OTHER GI SCALES IN SSc

Another scale that is not SSc-specific is the Gastrointestinal Quality of Life Index (64), which is a validated 52-item questionnaire capturing SSc-related gut dysfunction given to more than 400 SSc clinic patients assessing the frequency and impact of 5 categories of symptoms. There was a positive correlation between diarrhea scores and pulmonary fibrosis (r = 0.13), but not with other organs. In addition, limited cutaneous SSc and diffuse cutaneous SSc did not score differently; this is expected for GI disease in SSc, which is virtually universal (65).

GASTROESOPHAGEAL REFLUX DISEASE (GERD)

GERD is extremely common in SSc and is often severe. There are scales that have been used in SSc that assess GERD, such as the Frequency Scale for the Symptoms of GERD (FSSG), and visual analog scales (66). One study in severe GERD used gut pH measurements, which did not differentiate active treatment from placebo, but the study was negative with respect to the FSSG and quality of life (67). This study compared ranitidine to placebo on background double-dose proton pump inhibitors for severe GERD in SSc patients. Therefore, we cannot conclude if pH measurements of the gut are useful as an outcome in SSc randomized controlled trials. The testing is invasive and needs training to be performed. The trial was also likely underpowered.

DYSPNEA MEASUREMENTS USED IN SSc-ASSOCIATED INTERSTITIAL LUNG DISEASE (ILD)/PULMONARY FIBROSIS

There is no fully validated dyspnea questionnaire in SSc. In addition, the quality of life (QOL) in SSc patients with ILD may be impacted by cough, which is not captured on questionnaires that have been studied in SSc. Numerous dyspnea scales have been published in other diseases such as chronic obstructive pulmonary disease (COPD) and idiopathic pulmonary fibrosis.

The dyspnea questionnaire by Mahler et al includes the Baseline Dyspnea Index (BDI) and Transition Dyspnea Index (TDI) (68). The Modified Medical Research Council Scale and the Oxygen Cost Diagram are widely used tools for evaluation of limitation of activities due to dyspnea that are used in COPD (68–72) but not SSc-associated ILD. There is an activity of daily living dyspnea scale, the Modified Dyspnea Index, and dyspnea scales from the Scleroderma Health Assessment Questionnaire (HAQ) and the Symptom Burden Index (4,7,73). Dyspnea is significantly related to function and QOL. A model including age, sex, disease duration, disease severity, and dyspnea explained one-third, 10%, 40%, and 30%, respectively, of the variance of the HAQ (74). The BDI and TDI from the questionnaire by Mahler et al (68) will be reviewed more extensively since being used in a SSc lung disease trial of cyclophoshamide, and there has been some validation in SSc of these instruments (75). The Borg Dyspnea Index is a measurement by which dyspnea is assessed following the 6-minute walk and has only been partially validated in ILD and pulmonary hypertension (76). The Modified Borg Dyspnea Scale is numerical and describes the severity of dyspnea.

BASELINE DYSPNEA INDEX (BDI) AND TRANSITION DYSPNEA INDEX (TDI) (MAHLER'S INDEX)

Description

The BDI and TDI measure dyspnea at one point in time and then how it has changed at another time point (68). It can be self- or interviewer administered.

Purpose. To measure the severity of dyspnea at a point in time and in followup to determine if there is change (improving, the same, or worsening), as well as to evaluate the severity of dyspnea as the changes are added to the baseline score.

Content. For Mahler's dyspnea scales, the BDI is designed to rate the severity of dyspnea at a single time point, and the TDI is designed to capture a change (or no change) from the baseline assessment. Each index rates 3 different categories: magnitude of task, magnitude of effort, and functional impairment. Each category has 5 grades ranging from 0 (severe) to 4 (unimpaired) added together for a baseline focal score (range 0–12). At the transition period, changes in dyspnea were rated by 7 grades, ranging from −3 (major deterioration) to +3 (major improvement). The ratings for each of the 3 categories for the TDI were added to form a transition focal score (range −9 to +9) (68).

Developer. D. A. Mahler, D. H. Weinberg, C. K. Wells, and A. R. Feinstein.

Number of items in scale. Each index rates 3 different categories: magnitude of task, magnitude of effort, and functional impairment. Each category has 5 grades ranging from 0 (severe) to 4 (unimpaired). At the transition period, changes in dyspnea were rated by 7 grades, ranging from −3 (major deterioration) to +3 (major improvement).

Scoring. Each category is added together for a baseline score (range 0–12) as a maximum of 4 for each of 3 scales. For the TDI, the ratings for each of the 3 categories were added to form a transition focal score (range −9 to +9). TDI has improvement as major, moderate, or minor corresponding to improvement on the scale as 7–9, 4–6, and 1–3, respectively, and conversely there is deterioration if the scales show worsening (−1 to −3 for minor, −4 to −6 for moderate, and −7 to −9 for severe).

Reliability. This has not been fully tested in SSc, but the instrument was used successfully in a scleroderma lung study (SLS) using cyclophosphamide versus placebo (75).

Validity. The original indices were validated in men, most of whom had chronic obstructive pulmonary disease (68).

Construct validity. Not fully tested in SSc; in the SLS, baseline scores of the BDI and visual analog scale (VAS) for breathing were highly correlated (r = −0.61). Medical Outcomes Study Short Form 36 (SF-36) scores were able to differentiate patients with more breathlessness (measured by BDI and VAS for breathing) (77).

Face validity. In the SLS, there was face validity for the Mahler dyspnea scale, since a larger proportion of patients treated with cyclophosphamide obtained at least the minimum important difference (MID) compared to placebo in the TDI (78).

Predictive validity. Not tested in SSc.

Sensitivity to change. Using the SF-36 transition question and defining the MID as patients who rated themselves as a little better or a little worse in the SLS, the MID was estimated for the TDI. TDI improvement and worsening, respectively, ranged from 1.05 to 2.16 (mean 1.5) and from −0.61 to −2.55 (mean −1.5) (79). More patients on cyclophosphamide achieved a MID for the TDI (46% for cyclophosphamide versus 13% for placebo) (78). The mean TDI change was higher in the cyclophosphamide group (75). Other measurements such as changes in fibrosis on high-resolution computed tomography were associated with changes in dyspnea (80).

OTHER LUNG SCALES IN SSc

There is face validity of the Saint George's Respiratory Questionnaire (81) in the evaluation of the health-related quality of life in SSc associated with interstitial lung disease (82). An exercise program in SSc found that a significant proportion of patients with SSc experienced an improvement in the Saint George's Respiratory Questionnaire and exercise tolerance (83).

PULMONARY ARTERIAL HYPERTENSION (PAH) IN SSc

PAH randomized controlled trials often do not measure a dyspnea questionnaire. Therefore, standardized dyspnea questionnaires may or may not be sensitive to change in SSc-associated PAH. A Delphi exercise for PAH in SSc suggested the domains should include lung vascular, exercise testing, cardiac function, dyspnea (as measured by a visual analog scale [VAS]), discontinuation of treatment, quality of life, and physician global assessment. These could be measured by right heart catheterization, echocardiography, exercise oxygen saturation, 6-minute walk distance, Medical Outcomes Study Short Form 36, the Health Assessment Questionnaire (HAQ), and survival as well as adverse events (84).

The New York Heart Association and World Health Organization functional class systems are essentially the same and are divided into 4 categories: no restriction of activities (class I), mild restriction (class II), moderate (class III), and severe inability to do activities of daily living with dyspnea even at rest (class IV) (85,86). There is a large potential range of severity in class II and III patients, so refining a dyspnea questionnaire would be valuable. There is a lack of correlation between the HAQ in SSc and PAH with respect to functional class at baseline and with treatment (87). The SSc community via a Delphi exercise rated that outcome measurements in SSc PAH should include severity of dyspnea measured on a VAS (84).

PULMONARY ARTERIAL HYPERTENSION (PAH)–SPECIFIC QUALITY OF LIFE INSTRUMENT: CAMBRIDGE PULMONARY HYPERTENSION OUTCOME REVIEW (CAMPHOR)

Description

CAMPHOR is a PAH-specific quality of life (QOL) measure and not specifically for SSc. It is the first pulmonary hypertension–specific instrument for assessing patient-reported symptoms, functioning, and QOL, with scales including overall symptoms (made up of energy, breathlessness, and mood subscales), functioning and QOL. This has not been validated specifically in SSc-associated PAH (8).

Purpose. This instrument is to be used in PAH to assess QOL. It should quantify the effects of PAH on QOL, assessing impairment, disability, and needs-based QOL.

Content. Questions about symptoms, function, energy, mood, breathlessness, and QOL.

Developer. Galen Research (S. P. McKenna, N. Doughty, D. M. Meads, L. C. Doward, and J. Pepke-Zaba).

Number of items in scale. CAMPHOR has 3 scales including overall symptoms, functioning, and quality of life with 65 items (8). The overall symptoms category has the subscales of energy, breathlessness, and mood. The instrument consists of 25 items for impairment, 15 for functioning, and 25 for QOL.

Reliability. The CAMPHOR scales have good reproducibility (0.86–0.92 for test–retest correlations) when tested in idiopathic PAH (8), but it has not been tested in SSc.

Validity. The CAMPHOR scales have very good internal consistency ($\alpha = 0.90-0.92$) (7).

Face validity. The CAMPHOR utility score appears better able to distinguish between World Health Organization functional classes (II and III) than the EuroQol 5-domain and Short Form 6D (88).

Construct validity. The CAMPHOR scales have convergent, divergent, and known-groups validity (8).

Predictive validity. Patients remaining in the New York Heart Association (NYHA) class III experienced, on average, a significant improvement (CAMPHOR Utility Index and functioning), which exceeded the minimum important difference (MID) when PAH was treated (89).

Sensitivity to change. The CAMPHOR Utility Index has face validity and is responsive to change in PAH, but is not SSc specific. Patients can experience meaningful improvement even if they do not improve on functional class (which could require a larger change in status), and the CAMPHOR Utility Index distinguished between adjacent NYHA classes and correlated with 6-minute walking test (6MWT) results. CAMPHOR subscales and utility were as responsive as the 6MWT (effect sizes range 0.31–0.69 for the CAMPHOR and 0.16–0.34 for the 6MWT). The within-group MID for the CAMPHOR Utility Index is 0.09 (89).

Translations/adaptations. CAMPHOR has been validated in the US (90). It has also been adapted to French Canadian and English Canadian (91). There is also a version for English in Australia.

RAYNAUD'S PHENOMENON (RP) AND DIGITAL ULCERS

Raynaud's Condition Score has been validated in RP associated with SSc and will be discussed in detail, whereas there will only be a brief overview of DU.

DIGITAL ULCERS (DU)

It has been suggested that core measures for Raynaud's phenomenon (RP) in SSc clinical trials contain the Raynaud's Condition Score, physician and patient global assessments of RP activity, a DU measure, measures of disability and pain (Health Assessment Questionnaire), and measures of psychological function/generic quality of life (Arthritis Impact Measure 2 or Medical Outcomes Study Short Form 36) (6). Outcomes in DU are not standardized. There is no consensus on which DU measurements should be included in SSc DU trials. DU assessments may include visual analog scale (VAS) for RP, DU pain scales, number of digital ulcers, size of DU, burden of DU, healing or partial healing of DU (all or a cardinal ulcer), prevention of new DU, proportion of patients affected by DU, mean number of DU per patient, and VAS for physician and patient global assessments (92). As would be expected, within a 3-month SSc ulcer study there were significant correlations between ulcer dimension and pain VAS ($r = 0.42$, $P < 0.001$) (93).

RAYNAUD'S CONDITION SCORE (RCS)

Description

The RCS is a self-reported global assessment of Raynaud's phenomenon (RP) activity using a 0–10 ordinal scale, which incorporates the cumulative daily frequency, duration, severity, and impact of RP attacks. A composite score from daily measures is then calculated (6).

Purpose. To estimate the overall effect of RP.

Content. The RCS is a daily self-assessment of RP activity using a 0–10 ordinal scale that incorporates the cumulative daily frequency, duration, severity, and impact of RP attacks.

Developer. Peter Merkel, et al. Boston University.

Number of items in scale. One item with an 11-point ordinal scale (0–10), completed on a daily basis, and then an overall summary score is calculated for a defined period of time. There are no subscales.

Scoring. The number on the ordinal scale completed daily is added and divided by the number of days it has been completed to get a mean RCS for a period of time.

Reliability. It was found to be reliable when analyzing data from a trial (94).

Validity. In a recent randomized controlled trial (RCT) of tadalafil versus placebo, improvement in the frequency and duration of RP, RCS, healing and number of new digital ulcers (DUs), Scleroderma Health Assessment Questionnaire (SHAQ), and patient and physician global assessments significantly improved with active treatment (94).

Construct validity. Merkel et al have demonstrated the construct, content, criterion, and discriminant validity of the RCS, HAQ, and 12 visual analog scales (VAS) for RP in scleroderma using data from a RP RCT (6). There were relevant associations between the outcome measures and the patient and physician global assessments of RP activity.

Predictive validity. RCS can discriminate between those with and without DUs (6).

Sensitivity to change. RCS has been studied to determine the change needed to be clinically relevant in an RP trial. The minimum important difference score for the RCS for improvement is from −13.9 to −14.3 points (95). The patient acceptable symptom state was 34 (scale 0–100).

Critical Appraisal of Overall Value to the Rheumatology Community

Many clinicians and even researchers do not routinely use or interpret the RCS. In trials, it is often performed in addition to recording the frequency, severity, and duration of attacks. Therefore, the added value of the RCS is not fully determined. There is an advantage if a day of data are missing in an RP trial, since the score can still be calculated with the data that are completed, whereas if a day is missed then the frequency and duration of RP over 2 weeks cannot be calculated. There is a theoretical advantage to having a single scale that incorporates the impact of RP. Confusion between a 0–10 ordinal scale summary score of RCS and other Raynaud's scales may occur; however, the RCS is labeled accordingly whereas the other scales are often continuous VAS scores for RP.

HAND FUNCTION INSTRUMENTS FOR SSc

A detailed review for hand function scales has been performed elsewhere (96). However, some SSc studies related to hand function will be briefly reviewed. The Cochin Hand Function Scale has had good construct validity and its total score explained 75% of the variance of the Health Assessment Questionnaire in SSc patients (20). The Duruoz Hand Index was studied for test–retest reliability and intraclass correlation coefficients were very good (0.81–0.97) (48). The UK SSc Functional Score (17) and the Michigan Hand Questionnaire (97) also measure hand function. The latter may not be very useful for SSc hand function and has been rarely used for digital ulcer assessment (92).

THE MOUTH HANDICAP SCALE IN SSc

The majority of patients with SSc have oral problems including reduced oral opening, difficulties with dry mouth, and functional impairment with respect to oral hygiene. Mouthon et al have published the Mouth Handicap Scale in Systemic Sclerosis (98). It has 12 items with each scored from 0–4 (total score range 0–48). The mean ± SD total score of the scale was 20.3 ± 9.7. The test–retest reliability was 0.96. Divergent validity was confirmed for global disability (Health Assessment Questionnaire; $r = 0.33$), hand function (Cochin Hand Function Scale; $r = 0.37$), interincisor distance ($r = -0.34$), handicap (McMaster-Toronto Arthritis questionnaire; $r = 0.24$), depression and anxiety using the Hospital Anxiety and Depression ($r = 0.26$ and $r = 0.17$ for depression and anxiety, respectively). Three factors within the scale could explain two-thirds of the variance (98).

DISCUSSION

Many important instruments were not discussed in this review. Also, some articles may have been missed by the search strategy. Validation and reliability testing varied, where in some instruments (such as Raynaud's Control Score [RCS]) it was tested within a randomized controlled trial (RCT). For others there was cross-sectional testing at a single site (Symptom Burden Index). The University of California, Los Angeles, Scleroderma Clinical Trials Consortium Gastrointestinal Scale will likely be used within a treatment trial to determine its sensitivity to change. The Cambridge Pulmonary Hypertension Outcome Review needs validation in systemic sclerosis (SSc) if it is to be used for pulmonary arterial hypertension in SSc. However, for the instruments that were included, many have been partially validated in SSc, which is important for future research. Some lack testing for sensitivity to change. The global assessments (as in any rheumatic disease) do not have standardized questions or time frames but have been found to be sensitive to change within studies. There are also validated measurements that are not completed by the patients that are valuable in routine care and trials, such as the Modified Rodnan Skin Thickness Score.

There are also differences in minimum important differences (MIDs) when comparing how they were derived, such as in the Health Assessment Questionnaire (HAQ) (22,23,34). The MID in a trial of early diffuse cutaneous SSc (dcSSc) is not the same as expert determined. In the clinic with limited cutaneous SSc and dcSSc patients, many of whom may have been relatively stable, and in the latter methodology, the mean change in HAQ did not make sense since it was below the limit of the scale to detect change (22). This could also illustrate that patients may be worse with SSc that is unrelated to worsening function (and due to symptoms in other domains such as lung, gastrointestinal, Raynaud's phenomenon, etc.).

In addition, for use as outcomes in clinical trials, the sample size calculations can be different for instruments such as the HAQ, functional index, and physician global assessment due to variability in the measures in a group of SSc patients (99). This is important when selecting outcome measurements in clinical trials since some may be more apt to change within a given sample size.

AUTHOR CONTRIBUTIONS

Dr. Pope drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Khanna D, Lovell DJ, Giannini E, Clements PJ, Merkel PA, Seibold JR, et al. Development of a provisional core set of response measures for clinical trials of systemic sclerosis. Ann Rheum Dis 2008;67:703–9.
2. Khanna D, Distler O, Avouac J, Behrens F, Clements PJ, Denton C, et al. Measures of response in clinical trials of systemic sclerosis: the Combined Response Index for Systemic Sclerosis (CRISS) and Outcome Measures in Pulmonary Arterial Hypertension related to Systemic Sclerosis (EPOSS). J Rheumatol 2009;36:2356–61.
3. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
4. Steen VD, Medsger TA Jr. The value of the Health Assessment Questionnaire and special patient-generated scales to demonstrate change in systemic sclerosis patients over time. Arthritis Rheum 1997;40:1984–91.
5. Khanna D, Hays RD, Maranian P, Seibold JR, Impens A, Mayes MD, et al. Reliability and validity of the University of California, Los Angeles Scleroderma Clinical Trial Consortium Gastrointestinal Tract Instrument. Arthritis Rheum 2009;61:1257–63.
6. Merkel PA, Herlyn K, Martin RW, Anderson JJ, Mayes MD, Bell P, et al. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum 2002;46:2410–20.
7. Kallen MA, Mayes MD, Kriseman YL, de Achaval SB, Cox VL, Suarez-Almazor ME. The symptom burden index: development and initial findings from use with patients with systemic sclerosis. J Rheumatol 2010;37:1692–8.
8. McKenna SP, Doughty N, Meads DM, Doward LC, Pepke-Zaba J. The Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR): a measure of health-related quality of life and quality of life for patients with pulmonary hypertension. Quality Life Res 2006;15:103–15.
9. Johnson SR, Hawker GA, Davis AM. The Health Assessment Questionnaire disability index and Scleroderma Health Assessment Questionnaire in Scleroderma Trials: an evaluation of their measurement properties. Arthritis Rheum 2005;53:256–62.
10. Poole JL, Steen VD. The use of the Health Assessment Questionnaire (HAQ) to determine physical disability in systemic sclerosis. Arthritis Care Res 1991;4:27–31.
11. Clements PJ, Lachenbruch PA, Seibold JR, Zee B, Steen VD, Brennan P, et al. Skin thickness score in systemic sclerosis: an assessment of interobserver variability in 3 independent studies. J Rheumatol 1993; 20:1892–6.

12. Moser DK, Clements PJ, Brecht ML, Weiner SR. Predictors of psychosocial adjustment in systemic sclerosis: the influence of formal education level, functional ability, hardiness, uncertainty, and social support. Arthritis Rheum 1993;36:1398–405.
13. Poole JL, Williams CA, Bloch DA, Hollak B, Spitz P. Concurrent validity of the Health Assessment Questionnaire disability index in scleroderma. Arthritis Care Res 1995;8:189–93.
14. Malcarne VL, Greenbergs HL. Psychological adjustment to systemic sclerosis. Arthritis Care Res 1996;9:51–9.
15. Roca RP, Wigley FM, White B. Depressive symptoms associated with scleroderma. Arthritis Rheum 1996;39:1035–40.
16. Clements PJ, Wong WK, Hurwitz EL, Furst DE, Mayes M, White B, et al. Correlates of the disability index of the Health Assessment Questionnaire: a measure of functional impairment in systemic sclerosis. Arthritis Rheum 1999;42:2372–80.
17. Smyth AE, MacGregor AJ, Mukerjee D, Brough GM, Black CM, Denton CP. A cross-sectional comparison of three self-reported functional indices in scleroderma. Rheumatology (Oxford) 2003;42:732–8.
18. Sultan N, Pope JE, Clements PJ, Scleroderma Trials Study Group. The health assessment questionnaire (HAQ) is strongly predictive of good outcome in early diffuse scleroderma: results from an analysis of two randomized controlled trials in early diffuse scleroderma. Rheumatology (Oxford) 2004;43:472–8.
19. Khanna D, Furst DE, Clements PJ, Park GS, Hays RD, Yoon J, et al. Responsiveness of the SF-36 and the Health Assessment Questionnaire disability index in a systemic sclerosis clinical trial. J Rheumatol 2005; 32:832–40.
20. Rannou F, Poiraudeau S, Berezne A, Baubet T, Le-Guern V, Cabane J, et al. Assessing disability and quality of life in systemic sclerosis: construct validities of the Cochin Hand Function Scale, Health Assessment Questionnaire (HAQ), Systemic Sclerosis HAQ, and Medical Outcomes Study 36-Item Short Form Health Survey. Arthritis Rheum 2007;57:94–102.
21. Ouimet JM, Pope JE, Gutmanis I, Koval J. Work disability in scleroderma is greater than in rheumatoid arthritis and is predicted by high HAQ scores. Open Rheumatol J 2008;2:44–52.
22. Sekhon S, Pope J, Canadian Scleroderma Research Group (CSRG), Baron M. The minimally important difference (MID) for patient centered outcomes including Health Assessment Questionnaire (HAQ), fatigue, pain, sleep, global VAS and SF-36 in scleroderma (SSc). J Rheumatol 2010;37:591–8.
23. Khanna D, Furst DE, Hays RD, Park GS, Wong WK, Seibold JR, et al. Minimally important difference in diffuse systemic sclerosis: results from the D-penicillamine study. Ann Rheum Dis 2006;65:1325–9.
24. Schnitzer M, Hudson M, Baron M, Steele R. Disability in systemic sclerosis: a longitudinal observational study. J Rheumatol 2011;38:685–92.
25. Cole JC, Khanna D, Clements PJ, Seibold JR, Tashkin DP, Paulus HE, et al. Single-factor scoring validation for the Health Assessment Questionnaire-disability index (HAQ-DI) in patients with systemic sclerosis and comparison with early rheumatoid arthritis patients. Qual Life Res 2006;15:1383–94.
26. Johnson SR, Glaman DD, Schentag CT, Lee P. Quality of life and functional status in systemic sclerosis compared to other rheumatic diseases. J Rheumatol 2006;33:1117–22.
27. Georges C, Chassany O, Toledano C, Mouthon L, Tiev K, Meyer O, et al. Impact of pain in health related quality of life of patients with systemic sclerosis. Rheumatology (Oxford) 2006;45:1298–302.
28. Korn JH, Mayes M, Cerinic M, Rainisio M, Pope J, et al, for the RAPIDS-1 Study Group. Digital ulcers in systemic sclerosis: prevention by treatment with bosentan, an oral endothelin receptor antagonist. Arthritis Rheum 2004;50:3985–93.
29. Zelenietz C, Pope J. Differences in disability as measured by the Health Assessment Questionnaire (HAQ) between patients with and without digital ulcers in systemic sclerosis: a post hoc analysis of pooled data from two randomized controlled trials in digital ulcers using bosentan. Ann Rheum Dis 2010;69:2055–6.
30. Brower LM, Poole JL. Reliability and validity of the Duruoz Hand Index in persons with systemic sclerosis (scleroderma). Arthritis Rheum 2004;51:805–9.
31. Lawrence E, Pope J, Al Zahraly Z, Lalani S, Baron M. The relationship between changes in self-reported disability (measured by the Health Assessment Questionnaire-HAQ) in scleroderma and improvement of disease status in clinical practice. Clin Exp Rheumatol 2009;27 Suppl 54:32–7.
32. Khimdas S, Harding S, Bonner A, Canadian Scleroderma Research Group, Zummer B, Baron M, Pope J. Associations with digital ulcers (DU) in a large cohort of systemic sclerosis (SSc). Arthritis Care Res 2010;63:142–9.
33. Stevens A, Pope JE. Retest reliabilities and variability among scleroderma patients for 4 tests of disability: support for a better measure. Presented at the Canadian Rheumatology Association Meeting, 1995. J Rheumatol 1995;22:1603.
34. Gazi H, Pope JE, Clements P, Medsger TA, Martin RW, Merkel PA, et al. Outcome measurements in scleroderma: results from a Delphi exercise. J Rheumatol 2007;34:501–9.
35. Georges C, Chassany O, Mouthon L, Tiev K, Toledano C, Meyer O, et al. Validation of French version of the Scleroderma Health Assessment Questionnaire (SSc HAQ). Clin Rheumatol 2005;24:3–10.
36. Johnson SR, Baron M, Hudson M, Taillefer S, Hirsch A, and the Canadian Scleroderma Research Group. Lung VAS has incremental concurrent validity over the HAQ-DI was an outcome measure evaluating scleroderma lung disease. American College of Rheumatology Annual Meeting, San Diego, 2005. Arthritis Rheum 2005;52 Suppl:S584.
37. Kuwana M, Sato S, Kikuchi K, Kawaguchi Y, Fujisaku A, Misaki Y, et al. Evaluation of functional disability using the health assessment questionnaire in Japanese patients with systemic sclerosis. J Rheumatol 2003;30:1253–8.
38. Morita Y, Muro Y, Sugiura K, Tomita Y, Tamakoshi K. Results of the Health Assessment Questionnaire for Japanese patients with systemic sclerosis: measuring functional impairment in systemic sclerosis versus other connective tissue diseases. Clin Exp Rheumatol 2007;25:367–72.
39. La Montagna G, Cuomo G, Chiarolanza I, Ruocco L, Valentini G. HAQ-DI Italian version in systemic sclerosis. Reumatismo 2006;58: 112–5.
40. Khanna D, Clements PJ, Postlethwaite AE, Furst DE. Does incorporation of aids and devices make a difference in the score of the health assessment questionnaire-disability index? Analysis from a scleroderma clinical trial. J Rheumatol 2008;35:466–8.
41. Serednicka K, Smyth AE, Black CM, Denton CP. Using a self-reported functional score to assess disease progression in systemic sclerosis. Rheumatology (Oxford) 2007;46:1107–10.
42. Guillevin L, Ortonne JP. Treatment of scleroderma. Ann Med Interne (Paris) 1983;134:754–65. In French.
43. Ostojic P, Damjanov N. The Scleroderma Assessment Questionnaire (SAQ): a new self-assessment questionnaire for evaluation of disease status in patients with systemic sclerosis. Z Rheumatol 2006;65:168–75.
44. Ostojic P, Damjanov N. Indices of the Scleroderma Assessment Questionnaire (SAQ) can be used to demonstrate change in patients with systemic sclerosis over time. Joint Bone Spine 2008;75:286–90.
45. Sigl T, Ewert T, Stucki G. Patient-centered assessment of functional health in systemic sclerosis: where are we now? Z Rheumatol 2004;63: 463–9.
46. Silman A, Akesson A, Newman J, Henriksson H, Sandquist G, Nihill M, et al. Assessment of functional ability in patients with scleroderma: a proposed new disability assessment instrument. J Rheumatol 1998;25: 79–83.
47. Ruof J, Bruhlmann P, Michel BA, Stucki G. Development and validation of a self-administered systemic sclerosis questionnaire (SySQ). Rheumatology (Oxford) 1999;38:535–42.
48. Danieli E, Airo P, Bettoni L, Cinquini M, Antonioli CM, Cavazzana I, et al. Health-related quality of life measured by the Short Form 36 (SF-36) in systemic sclerosis: correlations with indexes of disease activity and severity, disability, and depressive symptoms. Clin Rheumatol 2005; 24:48–54.
49. Pope J, Baron M, Bellamy N, Campbell J, Carette S, Chalmers I, et al. The variability of skin scores and clinical measurements in scleroderma. J Rheumatol 1995;22:1271–6.
50. Fan X, Pope J, Canadian Scleroderma Research Group (CSRG), Baron M. What Is the relationship between disease activity, severity and damage in a large Canadian systemic cohort? Results from the Canadian Scleroderma Research Group (CSRG). Rheumatol Int 2010;30:1205.
51. Hudson M, Impens A, Baron M, Seibold JR, Thombs BD, Walker JG, et al. Discordance between patient and physician assessments of disease severity in systemic sclerosis. J Rheumatol 2010;37:2307–12.
52. Clements PJ, Hurwitz EL, Wong WK, Seibold JR, Mayes M, White B, et al. Skin thickness score as a predictor and correlate of outcome in systemic sclerosis: high-dose versus low-dose penicillamine trial. Arthritis Rheum 2000;43:2445–54.
53. Schieir O, Thombs BD, Hudson M, Boivin JF, Steele R, Bernatsky S, et al. Prevalence, severity, and clinical correlates of pain in patients with systemic sclerosis. Arthritis Care Res (Hoboken) 2010;62:409–17.
54. Malcarne VL, Hansdottir I, McKinney A, Upchurch R, Greenbergs HL, Henstorf GH, et al. Medical signs and symptoms associated with disability, pain, and psychosocial adjustment in systemic sclerosis. J Rheumatol 2007;34:359–67.
55. Benrud-Larson LM, Heinberg LJ, Boling C, Reed J, White B, Wigley FM, et al. Body image dissatisfaction among women with scleroderma: extent and relationship to psychosocial function. Health Psychol 2003; 22:130–9.
56. Thompson AE, Shea B, Welch V, Fenlon D, Pope JE. Calcium channel blockers for Raynaud's phenomenon in progressive systemic sclerosis. Arthritis Rheum 2001;44:1841–7.
57. Hudson M, Bernatsky S, Taillefer S, Fortin PR, Wither J. Patients with

systemic autoimmune diseases could not distinguish comorbidities from their index disease. J Clin Epidemiology 2008;61:654–62.

58. Sandusky SB, McGuire L, Smith MT, Wigley FM, Haythornthwaite JA. Fatigue: an overlooked determinant of physical function in scleroderma. Rheumatology (Oxford) 2009;48:165–9.
59. Hewlett S, Dures E, Almeida C. Measures of fatigue. Arthritis Care Res (Hoboken) 2011;63:S263–86.
60. Suarez-Almazor ME, Kallen MA, Roundtree AK, Mayes M. Disease and symptom burden in systemic sclerosis: a patient perspective. J Rheumatol 2007;34:1718–26.
61. Khanna D, Hays RD, Park GS, Braun-Moscovici Y, Mayes MD, McNearney TA, et al. Development of a preliminary scleroderma gastrointestinal tract 1.0 quality of life instrument. Arthritis Rheum 2007;57: 1280–6.
62. Frech T, Hays RD, Maranian P, Clements PJ, Furst DE, Khanna D. Prevalence and correlates of sleep disturbance in systemic sclerosis: results from the UCLA scleroderma quality of life study. Rheumatology (Oxford) 2011;50:1280–7.
63. Bodukam V, Hays RD, Maranian P, Furst DE, Seibold JR, Impens A, et al. Association of gastrointestinal involvement and depressive symptoms in patients with systemic sclerosis. Rheumatology (Oxford) 2011; 50:330–4.
64. Eypasch E, Williams JI, Wood-Dauphinee S, Ure BM, Schmulling C, Neugebauer E, et al. Gastrointestinal Quality of Life Index: development, validation and application of a new instrument. Br J Surg 1995; 82:216–22.
65. Thoua NM, Bunce C, Brough G, Forbes A, Emmanuel AV, Denton CP. Assessment of gastrointestinal symptoms in patients with systemic sclerosis in a UK tertiary referral centre. Rheumatology (Oxford) 2010; 49:1770–5.
66. Muro Y, Sugiura K, Nitta Y, Mitsuma T, Hoshino K, Usuda T, et al. Scoring of reflux symptoms associated with scleroderma and the usefulness of rabeprazole. Clin Exp Rheumatol 2009;27 Suppl 54:15–21.
67. Janiak P, Thumshirn M, Menne D, Fox M, Halim S, Fried M, et al. Clinical trial: the effects of adding ranitidine at night to twice daily omeprazole therapy on nocturnal acid breakthrough and acid reflux in patients with systemic sclerosis: -a randomized controlled, cross-over trial. Aliment Pharmacol Ther 2007;26:1259–65.
68. Mahler DA, Weinberg DH, Wells CK, Feinstein AR. The measurement of dyspnea: contents, interobserver agreement and physiologic correlates of two new clinical indexes. Chest 1984;85;751–8.
69. Mahler D, Wells C. Evaluation of clinical methods for rating dyspnea. Chest 1988;93:580–6.
70. Elliott MW, Adams L, Cockcroft A, Macrae KD, Murphy K, Guz A. The language of breathlessness. Am Rev Respir Dis 1991;144:826–32.
71. Stoller JK, Ferranti R, Feinstein AR. Further specification and evaluation of a new clinical index for dyspnea. Am Rev Respir Dis 1986;134: 1129–34.
72. Chhabra SK, Gupta AK, Khuma MZ. Evaluation of three scales of dyspnea in chronic obstructive pulmonary disease. Ann Thorac Med 2009;4:128–32.
73. Yoza Y, Ariyoshi K, Honda S, Taniguchi H, Senjyu H. Development of an activity of daily living scale for patients with COPD: the Activity of Daily Living Dyspnea scale. Nihon Kokyuki Gakkai Zasshi 2009;47: 858–64.
74. Baron M, Sutton E, Hudson M, Thombs B, Markland J, Pope J, et al. The relationship of dyspnoea to function and quality of life in systemic sclerosis. Ann Rheum Dis 2008;67:644–50.
75. Tashkin DP, Elashoff R, Clements PJ, Goldin J, Roth MD, Furst DE, et al. Cyclophosphamide versus placebo in scleroderma lung disease. N Engl J Med 2006;354:2655–66.
76. Borg G. Psychophysical basis of perceived exertion. Med Sci Sports Exerc 1982;14:377–81.
77. Khanna D, Clements PJ, Furst DE, Chon Y, Elashoff R, Roth MD, et al. Correlation of the degree of dyspnea with health-related quality of life, functional abilities, and diffusing capacity for carbon monoxide in patients with systemic sclerosis and active alveolitis: results from the Scleroderma Lung Study. Arthritis Rheum 2005;52:592–600.
78. Khanna D, Yan X, Tashkin DP, Furst DE, Elashoff R, Roth MD, et al. Impact of oral cyclophosphamide on health-related quality of life in patients with active scleroderma lung disease: results from the scleroderma lung study. Arthritis Rheum 2007;56:1676–84.
79. Khanna D, Tseng CH, Furst DE, Clements PJ, Elashoff R, Roth M, et al, for Scleroderma Lung Study Investigators. Minimally important differences in the Mahler's Transition Dyspnoea Index in a large randomized controlled trial: results from the Scleroderma Lung Study. Rheumatology (Oxford) 2009;48:1537–40.
80. Goldin J, Elashoff R, Kim HJ, Yan X, Lynch D, Strollo D, et al. Treatment of scleroderma-interstitial lung disease with cyclophosphamide is associated with less progressive fibrosis on serial thoracic high-resolution CT scan than placebo: findings from the scleroderma lung study. Chest 2009;136:1333–40.
81. Incalzi RA, Bellia V, Catalano F, Scichilone N, Imperiale C, Maggi S, et al. Evaluation of health outcomes in elderly patients with asthma and COPD using disease-specific and generic instruments: the Salute Respiratoria nell'Anziano (Sa.R.A.) Study. Chest 2001;120:734–42.
82. Beretta L, Santaniello A, Lemos A, Masciocchi M, Scorza R. Validity of the Saint George's Respiratory Questionnaire in the evaluation of the health-related quality of life in patients with interstitial lung disease secondary to systemic sclerosis. Rheumatology (Oxford). 2007;46:296–301.
83. Antonioli CM, Bua G, Frige A, Prandini K, Radici S, Scarsi M, et al. An individualized rehabilitation program in patients with systemic sclerosis may improve quality of life and hand mobility. Clin Rheumatol 2009;28:159–65.
84. Distler O, Behrens F, Pittrow D, Huscher D, Denton CP, Foeldvari I, et al. Defining appropriate outcome measures in pulmonary arterial hypertension related to systemic sclerosis: a Delphi consensus study with cluster analysis. Arthritis Rheum 2008;59:867–75.
85. Hatano S, Strasser T. Primary pulmonary hypertension. World Health Organization Geneva: WHO; 1975.
86. AHA medical/scientific statement. 1994 revisions to classification of functional capacity and objective assessment of patients with diseases of the heart. Circulation 1994;90:644–5.
87. Chow S, Pope JE, Mehta S. Lack of correlation of the health assessment questionnaire disability index with lung parameters in systemic sclerosis associated pulmonary arterial hypertension. Clin Exp Rheumatol 2008;26:1012–7.
88. McKenna SP, Ratcliffe J, Meads DM, Brazier JE. Development and validation of a preference based measure derived from the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses. Health Qual Life Outcomes 2008;6:65.
89. Meads DM, McKenna SP, Doughty N, Das C, Gin-Sing W, Langley J, et al. The responsiveness and validity of the CAMPHOR Utility Index. Eur Respir J 2008;32:1513–9.
90. Gomberg-Maitland M, Thenappan T, Rizvi K, Chandra S, Meads DM, McKenna SP. United States validation of the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR). J Heart Lung Transplant 2008;27:124–30.
91. Coffin D, Duval K, Martel S, Granton J, Lefebvre MC, Meads DM, et al. Adaptation of the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) into French-Canadian and English-Canadian. Can Respir J 2008;15:77–83.
92. Matucci-Cerinic M, Seibold JR. Digital ulcers and outcomes assessment in scleroderma [review]. Rheumatology (Oxford) 2008;47 Suppl 5:v46–7.
93. Toffolo SR, Furtado RN, Klein A, Watanabe S, Andrade LE, Natour J. Measurement of upper limb ulcers in patients with systemic sclerosis: reproducibility and correlation with pain, function, and quality of life. Nurs Res 2008;57:84–92.
94. Shenoy PD, Kumar S, Jha LK, Choudhary SK, Singh U, Misra R, Agarwal V. Efficacy of tadalafil in secondary Raynaud's phenomenon resistant to vasodilator therapy: a double-blind randomized cross-over trial. Rheumatology (Oxford) 2010;49:2420–8.
95. Khanna PP, Maranian P, Gregory J, Khanna D. The minimally important difference and patient acceptable symptom state for the Raynaud's condition score in patients with Raynaud's phenomenon in a large randomised controlled clinical trial. Ann Rheum Dis 2010;69:588–91.
96. Poole JL. Measures of hand function. Arthritis Care Res (Hoboken) 2011;63:S189–99.
97. Chung KC, Pillsbury MS, Walters MR, Hayward RA. Reliability and validity testing of the Michigan Hand Outcomes Questionnaire. J Hand Surg Am 1998;23:575–87.
98. Mouthon L, Rannou F, Berezne A, Pagnoux C, Arene JP, Fois E, et al. Development and validation of a scale for mouth handicap in systemic sclerosis: the Mouth Handicap in Systemic Sclerosis scale. Ann Rheum Dis 2007;66:1651–5.
99. Pope JE, Bellamy N. Sample size calculations in scleroderma: a rational approach to choosing outcome measurements in scleroderma trials. Clin Invest Med 1995;18:1–10.

Summary Table for Systemic Sclerosis Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
HAQ HAQ DI (3)	Measures function in 8 domains	Self-report	5 minutes to complete Questions not always relevant to patient, e.g., opening milk carton	Hand scoring; <1 minute	8 subscales scored each in worst answer from 0–3 Adds 1 point if aids used unless score already = 3 If 8 scales answered, then total score divided by 8 (divide total score by no. scales completed) 0 = no disability, 3 = worst score. Moderate impairment is often >1.0 Observation that adding aids in SSc; scleroderma may overestimate disability (40)	Reliability demonstrated in RP RCT (6) Good test–retest reliability when repeated after 2 days (33) ICCs range 0.38–0.76 for therapist watching activities and patient self-reporting (13)	Convergent/construct validity shown (17) Construct validity shown (10) Predictive validity shown in early dcSSc trials/clinical practice (18,31) Content validity not tested in SSc (9) Concurrent validity shown (13,30) Predictive low HAQ was associated with improved MD global 1 year later in clinical practice (31) A low HAQ predicted improved skin/patient global assessments in early dcSSc RCTs (1.5–5 times more likely to improve) (18) HAQ explained change in skin score in dcSSc RCT (R^2 = 0.57) (53) Face validity as mean HAQ worse in dcSSc than lcSSc subset (32) HAQ worsens over time in SSc in a year from 0.039 to 0.071, and in 3 years by ≥0.12 (24) HAQ and pain both related to SF-36 (PCS) (R = 0.70) (27) Structural validity between RA and SSc patients (25)	Responsiveness shown (4,6) MID in early dcSSc 0.10–0.14 (ES 0.15–0.21) (23) Clinical practice not very well proven For improvement was 0.0125; well below a change in measurement, but at 75th percentile it was 0.125, which is a minimum change in scoring of HAQ For worsening was 0.042 (below the minimal measurement change) and 0.217 for 75th percentile so MID for HAQ in clinical practice may be lacking (22) MID for HAQ 0.2–0.25 by MD Delphi consensus (34) HAQ related to work disability in SSc (21) If skin score improved then HAQ improved in active dcSSc RCT (ICC < 0.69) (19)	Measures function easily Inexpensive Compares to other diseases such as RA Used widely in RCTs and clinical practice Predicts mortality/morbidity and if very low HAQ at baseline in early dcSSc, predicts higher chance of improving skin score over next year (18)	Not predictive of all organ systems Not related to PAH changes (88) but dyspnea explained some variance of HAQ (52) Primarily measuring MSK-related function Mostly tested in dcSSc early trials Most patients in clinical practice worsen function so MIDs in clinical practice have wide range and may lack some face validity when change is less than the minimum change in measurement
SHAQ (4)	Measures function in 8 domains and separately VAS scales for GI, lung, vascular, digital ulcers and overall	Self-report	5 minutes to complete	1 minute Hand scoring	HAQ DI scored from 0–3 and 5 other VAS scales converted to 0–3 (GI, vascular, digital ulcers, lung, overall)	Yes (6)	Convergent construct (17) Construct not tested Predictive not tested Content validity, not tested Concurrent (6) VAS subscales significantly correlated with objective parameters (4)	Responsive (4,6) MIDs for VAS scales Not tested fully, but VAS for RP on SHAQ has ES 0.20 and SRM 0.15 (6)	HAQ DI for function and adds other areas of problems in RP (easy to measure and important to patients) Scope in SSc is more inclusive due to addition of subscales Fast, inexpensive, feasible	Not utilized as much as HAQ DI Some validity of the subscales is lacking Sometimes papers report the subscales scores added together and this is not the intended scoring of the SHAQ
Physician global assessment	Outcome assessor to estimate overall disease activity Also used for severity and damage	Physician-administered after reviewing patient and often labs/results of other organ function tests	Takes a few seconds to complete (but after a review of the patient's specific data)	Easy scoring (10-mm VAS or Likert change score)	Higher number is a worse score	Test–retest reliability; interobserver reliability tested (very good ICCs) (50) Agreement between MD and patient global assessments is ICC 0.38 (52)	Content reliability (50) SF-36 related to MD global (19)	MID by MD Delphi is 8–13 on 100-mm VAS for MD global (34) Responsive with other parameters: skin scores/breathlessness Low HAQ associated with improved MD global 1 year later in clinical practice (31) Improved MD global associated with skin score and HAQ (R^2 = 0.46) (53)	Gives overall number for SSc disease Sensitive to change and good face validity	No standardized question No standardization of measurement before global assessment completed Important to include both MD and patient global assessments since not strongly correlated and measure different perspectives
Patient global assessment	Overall assessment of SSc by patient	Self-report	<1 minute	Hand scored in <1 minute	Higher number is worse Score of 0 considered no activity (sometimes scale is overall where 100 = best and 0 = worst ever, but not usually)	Cannot test for interobserver variability as scale is completed by patient Test–retest reliability modest (33) Agreement between MD and patient global assessments ICC 0.38 (52)	Patient global improvement and skin score improvement related in 2 early dcSSc trials (r = ≥0.25) (18) SF-36 related to patient global (19)	MID tested in clinical practice for patient global is from 4 to 7 mm on 100-mm VAS (22) Responsive with other parameters: skin scores, breathlessness (52) MID by physician Delphi is 10–12 on 100-mm VAS, for patient global (34)	Patient reported overall SSc disease burden	Does not correlate strongly with MD global assessment Patient may be measuring overall disease burden (activity and damage or severity) and not disease activity alone Patients rate higher disease severity than physicians (52) Important to include both MD and patient globals as they are not very strongly correlated and thus measuring different perspectives
Patient VAS of SSc	Measures SSc pain	Self-report	≤1 minute	~2 seconds to hand score	Higher rating implies more pain Often 0–100-mm VAS or Likert scale	Test–retest very good (33)	Pain related to more RP, active ulcers, arthritis, GI symptoms. Worse pain with dsSSc (22) SSc pain higher than RA (in one study and not in another) (26,97,22) HAQ and pain both related to SF-36 (PCS) (R = 0.70) (27) Pain and fatigue related to physical function (59)	MID for pain in SSc clinical practice is 4–8 on 100-mm VAS (22) MID by expert opinion of pain is 0.2–0.3 out of 3 for pain VAS (34)	Easy to do/important to test pain in SSc as it is common and likely under-recognized	Pain can be attributed to many causes (skin with pruritis, ulcers, RP, GI, arthritis, etc.), so if wanting to intervene with pain, specific organ pain or symptoms may need to be asked Pain questionnaire is not standardized with respect to how it is asked
SBI (7)	Determine SSc QOL in 8 domains (skin, hand mobility, calcinosis, shortness of breath, eating, bowel, sleep, and pain) 5 questions per domain	Self-report	A couple of minutes	A few minutes	Questions answered in Likert scale Scoring for each domain 0–10. Number of domains with problems noted	Good reliability between items Distribution of scores ~equal between low/medium/high Patients had mean of 5.7 out of 8 problems	Significantly correlated with HAQ (0.3–0.6) Negatively correlated with SF-36 (PCS and MCS) Construct validity Focus groups tested to develop domains (61) Domains independent but significantly related except shortness of breath (between domain correlations low to moderate, desired for different domains)	Not tested	Easy to complete Fills gap for patient symptoms not fully included in other scales such as SHAQ, UCLA SCTC GIT Has some common domains that are not identical	Only tested in single site with 62 patients, more work on instrument needed Easy to complete Value-added above other QOL (generic) instruments/SSc tools Adding all 5 questions in each domain may not give comparable score between domains as questions are how much, how often, how much interference and thus a rare problem that interferes a lot May score the same as a common problem interferes a little

(continued)

Summary Table for Systemic Sclerosis Measures* *(Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
UCLA SCTC GIT 2.0 (5,62)	7 multi-item scales with areas of reflux, distention/ bloating, diarrhea, fecal soilage, constipation, emotional well-being, and social functioning For total score, 6 of 7 scales averaged (excluding constipation) 34-item revised instrument 2.0 revised from 52-item 1.0	Self-report	A couple of minutes	A minute	Items scored 0–3 with lower values showing better HRQOL Higher score indicates more GI problems Correlated with self-rated GI severity (r = 0.60) Discriminant validity Detects differences among no symptoms, very mild, mild, moderate, severe/ very severe (F = 31)	Overall, test–retest reliability ICC 0.81; coefficient α = 0.71 Internal consistency shown (Cronbach's α 0.67–0.91) (5) Test–retest reliability 0.68–0.89 (5) Severity scales/subscales (r = 0.2–0.5) and overall r = 0.6 Higher scores in subscales with MD diagnosis of each of GERD, gastroparesis, bacterial overgrowth, pseudo-obstruction, diarrhea, rectal incontinence, constipation	Face validity Symptoms correlated with GI diagnoses (5,62) Correlated with worse sleep and depressive symptoms (63,64) Social functioning/emotional well-being scales correlated with those of SF-36 (r = 0.36) Convergent and divergent validity shown (5)		Only SSc-specific GI scale validated for SSc No floor effect Small ceiling effect	May not be responsive in a GI SSc trial (ceiling effect) If symptoms are severe and improve by 30% they still may be considered frequent by the patient
RCS (6)	Daily self-assessment of RP activity using 0–10 ordinal scale that incorporates cumulative daily frequency, duration, severity, and impact of RP attacks Therefore in general a part of RP attack diary	Self-report	Moderate respondent burden as done in conjunction with recording daily attacks (frequency, severity, and duration) Diary must be completed daily	Scoring adds each day that diary is completed then divided by number of days recorded May take long time if RP trial is long, but is total number of attacks, duration, and severity that takes scoring time	Scored 0–10 ordinal scale Higher score is worse	Higher if digital ulcers vs no ulcers; patient globals did not differentiate (but MD globals did) (6)	Construct, content, criterion, and discriminant validity present (tested in RP trial) (6) Change associated with improved frequency and RP duration, improved digital ulcers (new and healing), SHAQ, and MD and patient global assessments. (95)	Sensitive to change (tested in RP trial) (6) MID for RCS improvement is 14 points on 100-mm VAS (96) ES 0.6 SRM 0.64 These are same or better than other RP VAS scales (6)	Used in clinical trials and supposed to take into account the entire RP impact for patient	Most clinicians and even researchers are not too certain about what the score means and what is clinically important regarding baseline score and change Still performed in addition to frequency, severity, and duration of attacks daily Confusion between 0–10 ordinal scale and 100-mm VAS may occur as RCS is an ordinal scale and overall severity or activity of RP is often measured as continuous VAS
CAMPHOR (7)	QOL scale for PAH Designed/validated in idiopathic PAH Overall symptoms (energy, breathlessness, mood subscales) Functioning and QOL	Self-report	10 minutes	At least a few minutes to score	Test–retest correlations 0.86–0.92 (7)	Good internal consistency (α = ≥0.9) (7)	Convergent, divergent validity Responsive to change Correlated with change in functional class (7,89,90)	Subscales and utility were as responsive as 6MWT (ES 0.31–0.69 for CAMPHOR and 0.16–0.34 for 6MWT) (90) MID for utility index estimated as ~0.09 (90)	PAH specific QOL scale Validated in idiopathic PAH, but not in SSc	Not tested in PAH of SSc but in idiopathic PAH and alone may be insufficient to reflect all QOL aspects of SSc as it is weighted especially towards the impact of dyspnea (not tight skin, pain, GI problems, arthritis, etc.).
BDI TDI (69)	Measure severity and change in dyspnea Each index rates 3 categories: magnitude of task, magnitude of effort, and functional impairment Each category has 5 grades from 0 (severe) to 4 (unimpaired) Categories added together for BDI (range 0–12) TDI has 7 grades from −3 (major deterioration) to 3 (major improvement) Ratings for TDI scales added for transition score (−9 to 0)	Self-report or interviewer-administered	10 minutes	A few minutes	12 = no impairment and 0 = the worst for BDI 9 is the best, 0 is no change and − is the worst for TDI TDI has improvement as major, moderate, or minor corresponding to improvement on scale as 7–9, 4–6, and 1–3 and conversely there is deterioration if worsening	Not fully tested	Face validity (correlated with VAS for breathing (r = −0.6) and pulmonary function (FVC and DL$_{CO}$) (78) Other validity not fully tested in SSc but related to other lung parameters in SSc lung study (79)	MID for TDI was 1.05–2.16 (mean 1.5) for improvement and from −0.61 to −2.55 (mean −1.5) for worsening. (80) RR >3 for achieving at least MID for transitional dyspnea index score in cyclophosphamide vs placebo in ILD (79)	Face validity and sensitive to change in scleroderma lung study (thus at least partially validated) (76)	Not proven for PAH in SSc No consensus on which dyspnea scale to use for ILD studies in SSc

* HAQ DI = Health Assessment Questionnaire disability index; SSc; scleroderma = systemic sclerosis; RP = Raynaud's phenomenon; RCT = randomized controlled trial; ICC = intraclass correlation coefficient; dcSSc = diffuse cutaneous SSc; MD = physician; lcSSc = limited cutaneous SSc; SF-36 = Medical Outcomes Study Short Form 36; PCS = physical component score; RA = rheumatoid arthritis; MID = minimal important difference; ES = effect size; PAH = pulmonary arterial hypertension; MSK = musculoskeletal; SHAQ = scleroderma HAQ; VAS = visual analog scale; GI = gastrointestinal; SRM = standardized response mean; SBI = Symptom Burden Index; QOL = quality of life; MCS = mental component score; UCLA SCTC GIT = University of California, Los Angeles, Scleroderma Clinical Trials Consortium Gastrointestinal Scale; HRQOL = health-related QOL; GERD = gastroesophageal reflux disease; RCS = Raynaud's Condition Score; CAMPHOR = Cambridge Pulmonary Hypertension Outcome Review; 6MWT = 6-Minute Walk Test; BDI = Baseline Dyspnea Index; TDI = Transition Dyspnea Index; FVC = forced vital capacity; DL$_{CO}$ = diffusing capacity for carbon monoxide; RR = relative risk; ILD = interstitial lung disease.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S118–S157
DOI 10.1002/acr.20532
© 2011, American College of Rheumatology

Measures of Adult and Juvenile Dermatomyositis, Polymyositis, and Inclusion Body Myositis

Physician and Patient/Parent Global Activity, Manual Muscle Testing (MMT), Health Assessment Questionnaire (HAQ)/Childhood Health Assessment Questionnaire (C-HAQ), Childhood Myositis Assessment Scale (CMAS), Myositis Disease Activity Assessment Tool (MDAAT), Disease Activity Score (DAS), Short Form 36 (SF-36), Child Health Questionnaire (CHQ), Physician Global Damage, Myositis Damage Index (MDI), Quantitative Muscle Testing (QMT), Myositis Functional Index-2 (FI-2), Myositis Activities Profile (MAP), Inclusion Body Myositis Functional Rating Scale (IBMFRS), Cutaneous Dermatomyositis Disease Area and Severity Index (CDASI), Cutaneous Assessment Tool (CAT), Dermatomyositis Skin Severity Index (DSSI), Skindex, and Dermatology Life Quality Index (DLQI)

LISA G. RIDER,[1] VICTORIA P. WERTH,[2] ADAM M. HUBER,[3] HELENE ALEXANDERSON,[4] ANAND PRAHALAD RAO,[5] NICOLINO RUPERTO,[5] LAURA HERBELIN,[6] RICHARD BAROHN,[6] DAVID ISENBERG,[7] AND FREDERICK W. MILLER[1]

INTRODUCTION

The idiopathic inflammatory myopathies, including adult and juvenile dermatomyositis (DM), polymyositis (PM), and inclusion body myositis (IBM), are rare systemic autoimmune diseases that are characterized by chronic proximal muscle inflammation and weakness. In previous decades, there were few commonly used outcome measures in myositis, and those outcome measures were not validated. Therefore, in the past the assessment of outcomes in therapeutic trials was focused on nonstandardized measurement of muscle strength and function only.

Over the last decade, however, 2 international collaborative groups, the International Myositis Assessment and Clinical Studies Group (IMACS) and the Paediatric Rheumatology International Trials Organisation (PRINTO), have defined consensus core set measures to assess myositis disease activity and damage in adults and children and have begun to validate and standardize these measures (1,2). IMACS and PRINTO have also developed preliminary definitions of improvement that can be used as outcomes for therapeutic trials. These response criteria combine the core set activity measures to determine clinically meaningful improvement (3,4). Our section on myositis assessment focuses first on these core set measures of disease activity, quality of life (which is part of the PRINTO core set of activity, but a separate assessment domain for IMACS), and disease damage. To date, most of the validation data available for these core set measures are in patients with juvenile DM, with more limited validation in adult patients with DM or PM. Despite these efforts, there are still important gaps in the validation of these core set measures, and no validation studies have yet been performed in patients with IBM, although they are now being used frequently in myositis therapeutic trials.

We end this article with tools that have been used primarily in research studies and a few therapeutic trials that

Supported in part by the Intramural Research Program of the NIH National Institute of Environmental Health Sciences.

[1]Lisa G. Rider, MD, Frederick W. Miller, MD, PhD: National Institute of Environmental Health Sciences, NIH, Bethesda, Maryland; [2]Victoria P. Werth, MD: Philadelphia VA Medical Center and University of Pennsylvania, Philadelphia; [3]Adam M. Huber, MD: IWK Health Centre and Dalhousie University, Halifax, Nova Scotia, Canada; [4]Helene Alexanderson, PhD, RPT: Karolinska Institute and Karolinska University Hospital, Solna, Stockholm, Sweden; [5]Anand Prahalad Rao, MD, Nicolino Ruperto, MD, MPH: Paediatric Rheumatology International Trials Organisation, IRCCS G. Gaslini, Genoa, Italy; [6]Laura Herbelin, BS, CCRP, Richard Barohn, MD: University of Kansas Medical Center, Kansas City; [7]David Isenberg, MD: University College London, London, UK.

Address correspondence to Lisa G. Rider, MD, Environmental Autoimmunity Group, Program of Clinical Research, National Institute of Environmental Health Sciences, NIH, CRC 4-2352, MSC 1301, 10 Center Drive, Bethesda, MD 20892-1301. E-mail: riderl@mail.nih.gov.

Submitted for publication January 31, 2011; accepted in revised form May 12, 2011.

have some supporting validation in certain subgroups of patients with myositis. These tools are primarily organ-specific measures, including strength and functional assessments and cutaneous assessment tools. Quantitative muscle testing and the IBM Functional Rating Scale are the most commonly used instruments to assess patients with IBM, and although they have little supporting validation in myositis, quantitative muscle testing has been well validated in other myopathies and has been used frequently as an end point in therapeutic trials for IBM.

Although the methods for the assessment of patients with myositis have been limited in their scope, great strides have been made in the last decade in the development of new partially validated tools and international multidisciplinary consensus in using these measures that should enhance our understanding of the diverse effects of myositis on many organ systems and the development of new therapies.

PHYSICIAN AND PATIENT/PARENT GLOBAL ACTIVITY

Description

Purpose. An overall rating of the disease activity related to myositis, defined as potentially reversible pathology or physiology resulting from the underlying disease process (1).

Content. The physician global assessment of disease activity is to be judged by the physician based on all of the information available at the time of the evaluation, including the subject's appearance, medical history, physical examination, laboratory testing, and prescribed medical therapy. Adult patients or parents of children with myositis completing the patient/parent assessments are asked to take into account all of the active inflammation in their own or their child's muscles, skin, joints, intestines, heart, lungs, or other parts of the body that can improve with treatment. Patients over 10 years of age might also be able to complete a global activity assessment independent of their parents' ratings (5). The global disease activity score is recorded on a 10-cm visual analog scale (VAS) that is often anchored at the end points and middle. For patients and parents, a smiley face is often included at the 0-cm anchor and a sad face at the 10-cm anchor to improve understanding of the scale. A 5-point Likert scale can also be used as an alternative to the VAS.

Number of items. 1 item, either a VAS or a Likert scale rating.

Response options/scale. For the VAS rating, a score of 0–10 (down to 1 decimal place) is used, and for the Likert scale, a grade of 0 (no disease activity), 1 (mild disease activity), 2 (moderate disease activity), 3 (severe disease activity), or 4 (extremely severe disease activity) is used. The 10-cm VAS may have better precision, sensitivity, and specificity, but the 2 scales correlate highly (5).

Recall period for items. Scoring of the global disease activity requires that the activity be assessed at present, although a recall period of up to 2–4 weeks for the components of global disease activity is acceptable for stable patients who are assessed less frequently.

Endorsements. The physician global disease activity has been included as a core set activity measure for patients with adult and juvenile polymyositis (PM), dermatomyositis (DM), and inclusion body myositis (IBM) by the International Myositis Assessment and Clinical Studies Group (IMACS) (5) and as a core set activity measure for juvenile DM by the American College of Rheumatology/Paediatric Rheumatology International Trials Organisation/European League Against Rheumatism (2). These measures are also part of the preliminary response criteria for adult and juvenile DM and PM (4,6).

Examples of use. This score is used in myositis therapeutic trials and is now part of the criteria for the preliminary definition of improvement in myositis (3) and natural history studies, particularly those validating new myositis assessment tools (2,7). Physician and patient/parent global activity assessments are also used as part of the preliminary response criteria for adult and juvenile DM and PM (3,4).

Practical Application

How to obtain. The physician and patient global activity assessment is available in publications using this as an assessment tool, free of charge (5). The IMACS web site (http://www.niehs.nih.gov/research/resources/collab/imacs/diseaseactivity.cfm) also hosts copies of these tools, including the grading scales and detailed instructions, along with example cases and sample ratings as training materials for physicians.

Method of administration. The physician global assessment is completed by the physician assessing the patient and includes factors in the subject's appearance, medical history, physical examination, laboratory testing, and physician's resultant medical therapy. The adult (or teenage) patient or parent of a juvenile patient completes the patient or parent global activity assessment during the clinic or study visit.

Scoring. A single score is derived by measuring the distance the vertical line is from the left-hand side of the horizontal VAS. The length of the VAS should also be measured, so that the score can be adjusted to a denominator of 10 cm. The Likert scale also results in a single score. Scoring takes <1 minute and is done by hand.

Score interpretation. Zero represents inactive disease, and higher scores represent more severe disease activity. From a study of 115 juvenile patients with idiopathic inflammatory myopathy (IIM) assessed by pediatric rheumatologists at baseline and at 4–6- and 7–9-month followup evaluations, a Likert scale score of 0 (inactive disease) corresponds to a VAS rating of 0.1 cm (95% confidence interval [95% CI] 0.0–0.2), a Likert scale rating of 1 (mild activity) corresponds to a VAS rating of 1.5 cm (95% CI 1.3–1.6), a Likert scale rating of 2 (moderate activity) corresponds to a VAS rating of 4.8 cm (95% CI 4.4–5.2), a Likert scale rating of 3 (severe activity) corresponds to a VAS rating of 7.6 cm (95% CI 7.0–8.2), and a Likert scale rating of 4 (extremely severe activity) corresponds to a VAS rating of 9.2 cm (95% CI 7.9–10.4) (5).

Respondent burden. The time to complete a global activity assessment is <1 minute.

Administrative burden. The time to complete the physician global activity assessment is <1 minute, but this requires integration with other assessment measures to derive an overall impression.

Translations/adaptations. The parent global activity has been used internationally in the native languages of the patients (2,8). Physician global activity has been studied and used in all subgroups of patients with myositis, including adult and juvenile PM, DM, and IBM. Patient or parent global activity has been used in juvenile and adult DM and PM patients. Global activity assessments have also been used in a number of other systemic rheumatic diseases.

Psychometric Information

Method of development. Physician and patient global activity assessments were first used in the assessment of and as core set activity measures and part of the response criteria for other systemic rheumatic diseases, including rheumatoid arthritis and juvenile idiopathic arthritis. They were then adopted and studied in myositis.

Acceptability. Missing data are not common, and floor and ceiling effects are not common. There can be measurement error if physicians and patients/parents do not look at their previous ratings as part of the determination of the current rating. Although the rating is based on a collection of objective data, it is somewhat subjective and based on the experience of the rater.

Reliability. *Internal consistency.* In terms of internal reliability, Spearman's correlation was excellent (Spearman's r = 0.89) for the correlation of the VAS to the Likert scale for physician global disease activity, and the intraclass correlation coefficient was 0.85 ($P < 0.0001$) (5).

Test–retest reliability. Not available.

Interrater reliability. In a study of pediatric rheumatologists assessing paper cases of juvenile DM, the κ coefficient for agreement in the Likert scale ratings of global disease activity was 0.88, and Cronbach's α was 0.99 (5). Physicians and patients or parents had relatively poor agreement between their ratings (weighted κ coefficients 0.33–0.34), whereas parents and teenage patients had relatively good interrater reliability (weighted κ coefficient 0.84) in a juvenile IIM natural history study (5).

Validity. *Content validity.* A group of pediatric rheumatologists reached consensus on 17 clinical parameters that they considered very or extremely important in the determination of juvenile DM global disease activity: 3 clinical parameters that were moderately important in their formulation of global disease activity and 9 variables that were unimportant to their rating of global disease activity (5).

Construct validity. Most studies validating other measures of disease activity have examined the construct validity of physician global activity with the measure whose validation was being tested, and those studies will be discussed below under each of the other measures. For adult PM/DM patients who were screened for therapeutic trials for refractory disease, physician global activity correlated best with serum muscle enzyme levels (Spearman's r = 0.6–0.7), whereas for juvenile IIM, physician global activity correlated best with extramuscular activity, muscle strength, and physical function assessed by the Childhood Myositis Assessment Scale (CMAS) and Childhood Health Assessment Questionnaire (C-HAQ; Spearman's r = 0.6–0.7) (8). Physicians' and parents' or patients' global activity score correlated moderately (Spearman's r = 0.37–0.63), whereas parents' and older juvenile patients' ratings correlated moderately to highly (Spearman's r = 0.57–0.84) in juvenile IIM patients (5). In a study of juvenile DM patients, the correlation of physician and parent global disease activity was moderate (Spearman's r = 0.57) (2). Parent global activity also correlated moderately with other core set measures of disease activity, including the CMAS, C-HAQ, Disease Activity Score, and physical summary score of the Childhood Health Questionnaire (Spearman's r = 0.42–0.65) (2).

Criterion validity. There is no gold standard upon which to assess criterion validity. Sometimes the physician global activity is used to assess criterion validity in studies validating other measures.

Ability to detect change. In a juvenile IIM natural history study of patients, the standardized response mean (SRM) for physician global activity was −0.71 for the Likert scale and −0.62 for the VAS at 4-month followup, and after 8 months was −0.58 for both scales. The SRM for parent global activity (−0.54) was similar to the physician global activity after 8 months (5).

For juvenile DM patients who were close to diagnosis or in need of new therapy, the SRM at 6-month followup was 1.6 (95% CI 1.4–1.8) for physician global activity and was 1.2 (95% CI 1.0–1.4) for parent global activity, both assessed on the VAS (2). Both physician and parent global activity ratings had good ability to discriminate between patients who improved and those who did not improve by physician or parent ratings of responses to therapy (2).

For treatment-refractory adult PM/DM patients enrolled in trials of cytotoxic agents, the overall SRM was −0.51, but was −1.5 for the group of patients who met criteria for response.

A group of adult and pediatric rheumatologists and neurologists reached consensus that, for patients with juvenile and adult PM/DM, the physician and patient/parent global activity score should improve by ≥20% to classify a patient as improved (6). An absolute value for the minimum clinically important difference has not been determined.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The data demonstrate that physician and patient or parent global activity scores are valid overall measures of disease activity, are considered integral in the evaluation of myositis patients, and are part of the core set of activity measures used by several international collaborative groups. The requirement that the patient be assessed by an experienced clinician reduces the likelihood of biases in reporting. The physician global activity score has good content validity and reliability, and both measures have good construct validity and excellent responsiveness in juvenile (ages 2–18 years) and adult PM/DM patients. The 2 measures are clearly distinct.

Caveats and cautions. To reduce variability, this measure requires training of the person performing the assessment. The VAS may be slightly subjective and somewhat dependent on the experience of the rater. Neither physician nor patient global activity assessments have been formally validated in IBM.

Clinical usability. The measure should be useful in the assessment of myositis patients, particularly for longitudinal monitoring. Looking at previous measurements in formulating serial ratings is helpful to reduce measurement error. Patients ages >10 years may complete a global activity assessment.

Research usability. Both physician and patient/parent global activity assessments are well suited to use in research and are becoming widely used in myositis studies and therapeutic trials. They are considered to be a core assessment of disease activity.

MANUAL MUSCLE TESTING (MMT)

Description

Purpose. To measure muscle strength as part of the physical examination. No additional equipment is needed. MMT has been widely used in myositis therapeutic trials and clinical studies, previously as a primary end point (1) and more recently as part of a composite end point of core set measures (3). MMT has been reported most often as a summary score of a total number of proximal, distal, and axial muscle groups tested bilaterally or as a proximal score that sums a number of proximal muscle groups from the upper and lower extremities (1,6). More recently, MMT has been modified to a shorter version called MMT8, in which 8 proximal, distal, and axial muscle groups tested unilaterally closely approximate a total MMT score of 26 muscle groups tested bilaterally (9).

Content. Both the modified Medical Research Council (MRC) muscle strength scale and the Kendall grading scale are used (1,6). The modified 0–10-point Kendall grading scale provides firm definitions, along with plus (+) and minus (−) grades that provide an expanded scale. Muscle groups typically chosen include a combination of proximal, distal, and axial muscle groups.

Number of items. In inclusion body myositis (IBM) studies, 28 muscle groups are usually studied bilaterally, including shoulder abduction, elbow flexion and extension, wrist flexion and extension, hip flexion and extension, knee flexion and extension, ankle dorsiflexion and plantar flexion, and hip abduction. Neck flexion and extension are also tested (10). In polymyositis (PM), dermatomyositis (DM), and juvenile DM, 26 muscle groups are frequently tested (1,6) and include the above-listed muscle groups except for elbow extensors, but often there is no standardization in the number of muscle groups used. In some trials, only proximal MMT scores are reported, as proximal muscle groups are more affected than distal muscles in PM and DM (9). Recently, a subset of 8 muscle groups, including the neck flexors, deltoids, biceps, wrist extensors, gluteus maximus and medius, quadriceps, and ankle dorsiflexors, tested unilaterally was shown to have similar validity as the total MMT score; other sets of 8 proximal, distal, and axial muscle groups also performed equally well (11).

Response options/scale. The MRC grades were as follows: 0 = no contraction, 1 = flicker or trace of contraction, 2 = active movement with gravity eliminated, 3 = active movement against gravity, 4 = active movement against gravity and resistance, and 5 = normal power. This scale has been expanded to a 10-point scale in which the ability to resist against varying degrees of pressure in the antigravity position or the ability to move through varying ranges of motion in the gravity-eliminated position earns either a plus (+) or minus (−) in association with a particular grade. The Kendall 0–10-point scale similarly provides an expanded scale by assigning grades to hold the test position against varying degrees of pressure in the gravity-eliminated position or grading the ability to move through full or partial range of motion in the gravity-eliminated position (6).

Recall period for items. Scoring MMT requires that the activity be performed at the time MMT is administered (i.e., no recall period).

Endorsements. MMT has been included as a core set activity measure for adult and juvenile PM, DM, and IBM by the International Myositis Assessment and Clinical Studies Group (IMACS) (1) and as a core set activity measure for juvenile DM by the Paediatric Rheumatology International Trials Organisation/American College of Rheumatology/European League Against Rheumatism (2). Muscle strength testing, as assessed by MMT, is also part of the preliminary response criteria for adult and juvenile DM and PM (4,6).

Examples of use. Myositis therapeutic trials (1,6) and natural history studies (12).

Practical Application

How to obtain. MMT is available in publications that have used it as an assessment tool, free of charge (6). The IMACS web site (http://www.niehs.nih.gov/research/resources/collab/imacs/diseaseactivity.cfm) hosts a number of materials about MMT, including the grading scales, detailed instructions, and training videos.

Method of administration. MMT is administered by a trained therapist or clinician while observing the patient.

Scoring. Each muscle group tested is scored by using either the modified MRC or Kendall grading scale, depending on how much the muscle group can do in terms of moving against gravity or against applied pressure. Scores for individual muscle groups range from 0–5 on the MRC scale or from 0–10 on the Kendall scale, which are ordinal grading scales. The scores are summed for a total score or for subscores involving particular muscle groups (proximal, distal, axial scores). Computer programming is not necessary. Missing muscle groups are deleted from the value of the denominator, and the total score is adjusted to the new denominator, so that the percentage of maximum can be obtained.

Score interpretation. Using the Kendall 0–10 scale, the total MMT score ranges from 0–260 when 12 muscle groups are tested bilaterally along with 2 axial muscle groups. A proximal score of 0–160 represents 8 muscle groups tested bilaterally, a distal score of 0–80 includes 4 muscle groups tested bilaterally on the Kendall scale, and an axial score of 0–20 tests neck flexors and extensors. Normal strength is represented by a higher score at or near the top of the scale. The following interpretations of the scores of individual muscle groups have been used by researchers using MMT to study myositis: a muscle group graded from 0–3 on the Kendall scale indicates severe weakness, grade 4–6 indicates moderate weakness, grade 7–9 indicates mild weakness, and grade 10 indicates no detectable weakness (9). Validated cut points and the clinical meaning of MMT scores have not been established.

Respondent burden. If all items are attempted, the MMT can take 30–60 minutes to test 26–28 muscle groups. For the MMT8, testing takes <5 minutes. For the weak patient, the testing can be physically demanding and fatiguing, and in our clinical experience, it is important to adequately rest such patients before performing the test.

Administrative burden. The time it takes to administer the full MMT may be a limitation in a busy clinic, and such testing is typically assigned to a physical therapist to perform in a separate session. Scoring takes <1 minute and can be done by hand. Training in the administration of MMT is important, and can be obtained in local physical therapy or rehabilitation medicine departments. Contributions to measurement error can include inexperience of the examiner, improper positioning of the patient, bias in the application of force or in grading, and inconsistent commands (6). Rheumatologists, for example, typically score patients higher than experienced physical therapists.

Translations/adaptations. MMT is used internationally. It has been studied and used in all subgroups of myositis, including adult and juvenile PM, DM, and IBM. MMT has been used to assess strength in a variety of neuromuscular conditions.

Psychometric Information

Method of development. The MRC scale was developed by British physicians during World War II to grade strength after injuries. It was expanded and adapted to neuromuscular research in the 1970s. The shift from the MRC scale to the Kendall grading scale occurred in therapeutic trials of PM/DM in the 1990s because researchers sought to increase the sensitivity and specificity of MMT by expanding the grading scale with clear definitions. MMT had been in widespread use in therapeutic trials but has been validated for myositis only recently. The MMT8 was developed recently as a short form of the MMT that could be more practically applied by physicians testing patients in the clinical setting.

Acceptability. Although the tool is administered by the therapist or clinician, missing data can be common due to injury or joint contracture. If the data are absent due to an injury, they can be treated as an intent-to-treat point. There are recognized ceiling effects, particularly with known insensitivity of MMT for grades >3 of 5, where variations in the weight of patients' extremities and in the force applied by the examiner can result in discrepancies in detecting mild weakness.

Reliability. *Internal consistency.* In terms of internal reliability, Spearman's correlation was excellent for proximal MMT and MMT8 scores compared with the total MMT score in patients with treatment-refractory adult PM/DM (Spearman's r = 0.91–0.96) and 73 juvenile idiopathic inflammatory myopathy (IIM) patients from a natural history study (Spearman's r = 0.96–0.98) (11). Internal consistency, measured by Cronbach's α, was also very good to excellent for total, proximal, and MMT8 scores, ranging from 0.79–0.93 in adult PM/DM and from 0.90–0.93 in juvenile IIM (11).

Test–retest reliability. In a study of juvenile IIM patients who were evaluated by 1 pediatric physical therapist in the morning and again in the afternoon, the Spearman's rank correlation coefficient for the total, proximal, distal, and MMT8 scores for each pair ranged from 0.8–0.95 (all $P < 0.001$) (11). For individual muscle groups, the Spearman's rank correlation ranged from 0.70–1.0 (all $P < 0.04$) (13).

Interrater reliability. In a study of juvenile DM patients, interrater reliability was very good, with Kendall's W ranging from 0.71–0.76 for total, proximal, and MMT8 scores (11,13). The distal score had lower reliability (Kendall's W = 0.51) (13). The reliability of individual muscle groups varies and can be quite poor in distal and upper extremity proximal muscle groups (Kendall's W = 0.04–0.76) (13); therefore, it is important to use summary scores, particularly in research studies.

In a study of adult PM/DM patients, the interrater reliability (assessed by an intraclass correlation coefficient >0.65 or a ratio of the estimates of the standard error attributable to the physicians to the standard error attributable to the patients <0.4), was good for deltoids, biceps, quadriceps, gluteus medius and maximus, and ankle, and was poor for the neck flexors and wrist extensors (14).

Validity. *Content validity.* In developing the MMT8, a group of adult and pediatric rheumatologists and physical therapists agreed upon 3 possible combinations of 8 proximal, distal, and axial muscle groups that closely approximate a total MMT score and could be used in the clinic or in research settings for patients with juvenile and adult DM and PM (11).

Construct validity. In patients with juvenile PM/DM, total MMT, proximal MMT, and MMT8 scores correlated highly with physical function assessed by the Childhood Myositis Assessment Scale (Spearman's r = 0.70–0.73), and moderately with physician global activity (Spearman's r =0.49–0.54), functional disability measured by the Childhood Health Assessment Questionnaire (Spearman's r = 0.59–0.64), and magnetic resonance imaging (MRI), a score reflecting an average of activity and damage (Spearman's r = 0.45–0.48). MMT scores did not correlate significantly with serum muscle enzymes in patients with juvenile IIM (11).

In patients with adult PM/DM, total MMT, proximal MMT, and MMT8 scores correlated moderately with physical function measured by the Convery Activities of Daily Living Scale (Spearman's r = 0.59–0.70) and MRI (Spearman's r = 0.43–0.50). Correlations with physician global activity (Spearman's r = 0.33–0.37) and creatine kinase (Spearman's r = 0.34–0.38) were mild but significant (11).

Criterion validity. There is no gold standard upon which to assess criterion validity.

Ability to detect change. The standardized response mean (SRM) for total MMT was 0.56 in patients with juvenile PM/DM and 0.75 in patients with adult PM/DM in patients reassessed 4 months after baseline evaluation (11). The relative efficiency for proximal MMT (relative to the SRM for the total MMT score) was 0.98 in juvenile DM and 1.08 in adult PM/DM, and for the top MMT8 score was 1.16 in juvenile PM/DM and 1.24 for adult PM/DM (11).

In a study of juvenile DM patients enrolled at diagnosis or requiring escalation of therapy and assessed 6 months later, the SRM for total MMT was 1.2 (95% confidence interval 0.9–1.4) (2). Total MMT was also noted to have good discriminant validity (2).

A group of adult and pediatric rheumatologists and neurologists has reached consensus that MMT should improve by ≥15% to classify an adult PM/DM patient as improved and should improve by ≥18% to classify a juvenile PM/DM patient as improved (6). An absolute value for the minimum clinically important difference has not been determined.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The data demonstrate that MMT is a valid measure of strength, which is considered an integral assessment in the evaluation of myositis patients and part of the core set of activity measures by several international collaborative groups. The requirement that the patient is assessed by an experienced clinician reduces the likelihood of biases in reporting. MMT does not require any special equipment, except for a plinth or table on which the subject can lie flat. MMT has good to excellent reliability when used as a score that sums a number of muscle groups. It has good construct validity and excellent responsiveness in juvenile (ages 4–18 years) and adult PM/DM. It is also widely used to assess patients with IBM.

Caveats and cautions. To be performed appropriately and to reduce variability, training is required of the person performing the test. Subjects will need to be placed in positions that will be difficult for them to achieve as their weakness progresses. MMT also has decreased sensitivity and specificity in detecting mild weakness. The total MMT takes a long time to administer, but the MMT8, a subset of 8 muscle groups that performs similarly to the total score, is a good substitute in the busy clinical setting. MMT cannot reliably be used to assess children ages <5 years who have limited ability to cooperate. Like other measures of strength and function, MMT does not discriminate between activity and damage and may diminish in sensitivity and specificity as an activity measure for patients who are farther along in their illness course with accumulated damage and progressive muscle atrophy. MMT is frequently used but has not been formally validated in IBM.

Clinical usability. For some clinicians, the time required for administration limits the usefulness of MMT in the clinical context; however, the MMT8 is more usable in the clinical setting. Many clinicians have found MMT extremely useful for longitudinal monitoring of myositis patients.

Research usability. MMT is well suited to use in research and has been widely used in myositis studies. Concerns about ceiling effects may mean that it should be used with caution in patients with milder disease and that it will not be sensitive to change in patients with longstanding disease and a lot of muscle atrophy. Resources need to be invested to train a health care provider to perform these studies for a clinical trial.

HEALTH ASSESSMENT QUESTIONNAIRE (HAQ)/CHILDHOOD HEALTH ASSESSMENT QUESTIONNAIRE (C-HAQ)

Description

Purpose. The Stanford HAQ is a brief self-report questionnaire assessing physical function pertaining to activities of daily living in a variety of domains (15). Originally developed for use in rheumatoid arthritis, it has been successfully applied to a variety of rheumatic conditions, including idiopathic inflammatory myopathy (IIM) (16,17).

A modified version of the HAQ has been used that includes a variety of transitional questions intended to improve the responsiveness of the original tool (18). Although the modified HAQ has been used in myositis (19), there are little specific data regarding its psychometric properties in myositis.

The C-HAQ was adapted directly from the HAQ and was first published in 1994 (20). It was initially used in children with arthritis, but subsequently it has been evaluated in a variety of pediatric illnesses, including juvenile IIM (21,22). Its brevity and simplicity make it useful for longitudinal monitoring of children with juvenile IIM in the clinic setting.

General information on the HAQ and C-HAQ is covered in several other articles in this issue, specifically the article on measures of functional status and quality of life in rheumatoid arthritis, and only myositis-specific information is discussed here.

Endorsements. The HAQ has been included as a core set measure by the International Myositis Assessment and Clinical Studies Group (IMACS) (1), and the C-HAQ has been endorsed as a core set activity measure by both IMACS (1) and the Paediatric Rheumatology International Trials Organisation for juvenile IIM (8). These instruments are also part of the preliminary response criteria for adult and juvenile dermatomyositis (DM) and polymyositis (PM) (4,6).

Examples of use. The HAQ and C-HAQ have been used as part of myositis natural history studies, and recently have been incorporated as measures of physical function in myositis therapeutic trials (23–25).

Practical Application

How to obtain. The HAQ and C-HAQ are available from the original publications free of charge (15,20). They are also available from a variety of internet sites, including the IMACS web site: http://www.niehs.nih.gov/research/resources/collab/imacs/diseaseactivity.cfm.

Psychometric Information

Acceptability. The HAQ and C-HAQ are brief, and the language is generally at an appropriate level. It is not uncommon for respondents to neglect to complete the sections on the use of aids or assistance to complete tasks. It is recognized that the HAQ and C-HAQ have significant floor effects in all applications (patients with no or mild physical dysfunction cluster near 0).

Reliability. *Internal consistency.* The HAQ has not been formally assessed in adult IIM. In juvenile IIM patients, item-total correlations ranged from 0.35–0.81 by Spearman's r (all $P < 0.0001$), with only 4 items with a Spearman's r < 0.50 (21). Each domain of the C-HAQ also correlated well with the total score (Spearman's r = 0.59–0.84) (21).

Test–retest reliability. The HAQ has not been formally assessed in adult IIM. The intraclass correlation coefficient (ICC) was 0.87 for a group of juvenile IIM patients with <10% change by sphygmomanometry of the left hip abductor on consecutive visits (22). For patients with <10% change in the visual analog scale (VAS) of overall illness severity, the ICC was 0.96 (22).

Intra- and interrater reliability. Not applicable.

Validity. *Content validity.* The HAQ and C-HAQ have not undergone assessment of content validity in adult or juvenile IIM.

Construct validity. The HAQ has not been formally evaluated. However, in a longitudinal cohort study of patients with PM, DM, or overlap myositis, muscle strength measured by Manual Muscle Testing (MMT) correlated moderately with the HAQ (r = −0.61, $P < 0.0001$) and mildly but significantly with physician global disease activity (r = 0.28, $P = 0.009$) (23). The HAQ also correlated moderately with subscales of the Short Form 36, including the physical function, role function, body pain, and role emotional domains (r = 0.42–0.71) (23). In a study validating the Myositis Activities Profile (MAP) for adult PM/DM, a tool to assess limitations of physical activities in IIM patients, the HAQ had a Spearman's r of 0.70 with the MAP (24). In a study of adult PM/DM, the HAQ was shown to correlate significantly with muscle strength testing on the Medical Research Council grading scale and with the Henriksson and Sandstedt measure of functional disability ($P < 0.01$, correlation not provided), but not with isokinetic muscle strength testing (19). The HAQ correlated mildly but significantly with patient global activity (Pearson's r = 0.34).

In patients with juvenile IIM, the C-HAQ correlated moderately with physician global illness severity VAS (Spearman's r = 0.71, $P < 0.002$) and with hip abduction and shoulder abduction sphygmomanometry (Spearman's r = −0.57, $P < 0.002$ and Spearman's r = −0.51, $P < 0.01$, respectively) (22). Correlations were lower for knee extension and grip strength sphygmomanometry (Spearman's r = −0.40, $P = 0.05$ and Spearman's r = −0.079, $P > 0.20$, respectively), as expected (22).

In juvenile IIM, the C-HAQ correlated strongly (Spearman's r = >0.7) with the Childhood Myositis Assessment Scale; moderately (Spearman's r = 0.4−0.7) with physician global disease activity and physician global skin disease activity (by 10-cm VAS), MMT, Steinbrocker functional class, VAS for patient/parent global overall health, illness severity, and muscle symptoms, and the Disease Activity Score; and weakly (Spearman's r = <0.4) with physician global disease damage and skin disease damage (2,21). In another study of juvenile IIM patients, the C-HAQ showed good correlations with handheld dynometric muscle strength testing (partial correlation adjusted for age = −0.72, $P < 0.01$) (25). In a study of magnetic resonance imaging in juvenile DM, the C-HAQ correlated well with T2 relaxation time (Pearson's r = 0.49−0.58, $P < 0.001$) (26).

The C-HAQ correlated moderately with the total and muscle severity scores of the Myositis Damage Index (Spearman's r = 0.45–0.48, $P < 0.0001$) in juvenile IIM patients with a median disease duration of 6.8 years (27).

Criterion validity. Although not formally assessed for criterion validity, HAQ scores increase over time in cohort studies of adult DM and PM patients (16,23). HAQ scores are higher in patients who previously developed avascular necrosis or a compression fracture (16) and in patients with a chronic continuous or polycyclic illness course, osteoporosis, or who have a longer disease duration (23).

Ability to detect change. Data are not available for the HAQ. However, for the C-HAQ, in juvenile IIM patients enrolled at diagnosis, the responsiveness coefficient was 0.90 (22). For juvenile IIM patients with an improvement of >1 cm on the 10-cm VAS for physician global disease activity over 2 evaluations spanned by 7–9 months, the standardized response mean (SRM) and effect size were 0.87 and 0.67, respectively (21). The C-HAQ showed an SRM of 1.3 in juvenile IIM patients judged by the treating physician to have improved over 6 months (2). The change in C-HAQ scores correlates highly with change in the physical summary score of the Childhood Health Questionnaire (r = −0.73) (28).

A group of adult and pediatric rheumatologists and neurologists has reached consensus that physical function should improve by ≥15% to classify a patient as improved for patients with adult and juvenile PM/DM (6). An absolute value for the minimum clinically important difference has not been determined.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The HAQ and C-HAQ measure physical function, a domain of considerable importance to IIM patients and their health care providers. The tools are brief and take little time, no equipment, and minimal training to administer. They can be used in a variety of contexts (clinic, mail, internet, or phone) and are available in a variety of languages (29). They have been used extensively for a variety of illnesses. Finally, given that they are completed by the patient, parent, or caregiver, they have the advantage of being patient oriented. The C-HAQ has good reliability and excellent construct validity and responsiveness in patients with juvenile (ages 2–18 years) and adult PM/DM.

Caveats and cautions. From a development point of view, it is not clear that the HAQ or C-HAQ has undergone rigorous attempts to ensure content validity in patients with adult and juvenile IIM. Like other measures of strength and function, the HAQ and C-HAQ do not discriminate between activity and damage and may have poor sensitivity and specificity as a measure of activity for patients with moderate to severe damage, including patients who have muscle atrophy and fixed joint contractures. From an interpretation point of view, the biggest problem with the HAQ and C-HAQ is the floor effect. As patients improve and approach mild physical dysfunction, scores cluster near 0, and there is little room to document further improvement (21). The C-HAQ has been extensively validated in juvenile IIM. Data on validation of the HAQ in adult patients with PM/DM are incomplete, mainly confined to limited construct and criterion validity, and the HAQ has not been studied in inclusion body myositis.

Clinical usability. There are limited data to support the use of the HAQ in IIM, particularly to assess disease activity, although it still may be useful. The C-HAQ appears to have good reliability, validity, and responsiveness, making it a useful aid in guiding clinical decisions. Its simplicity, brevity, and ease of scoring minimize both administrative and respondent burden, facilitating its routine use in the clinic.

Research usability. There are limited data on construct validity and criterion validity for the HAQ in adult PM/DM, although it still may be useful. The documented reliability, validity, and responsiveness of the C-HAQ support its use in research. As in the clinical situation, its simplicity, brevity, and ease of scoring minimize the use of research resources. As noted, the floor effect may limit its usefulness in some research (e.g., involving patients with milder or more chronic disease).

CHILDHOOD MYOSITIS ASSESSMENT SCALE (CMAS)

Description

Purpose. The CMAS is an observational performance-based instrument that was developed to evaluate muscle strength, physical function, and endurance in children with juvenile idiopathic inflammatory myopathy (IIM) (30,31). First published in 1999, it has not been revised or updated.

Content. Items of the CMAS were chosen to explicitly include upper and lower extremity muscle groups, simple and compound movements, and timed items to evaluate endurance. The tool is purposefully weighted toward lower extremity proximal and axial muscle groups more than upper extremity and distal muscle groups to reflect the pattern of weakness in juvenile myositis (9). The tool is not divided into specific domains.

Number of items. The CMAS consists of 14 items, with no subscales.

Response options/scale. Specific scoring options are provided for each item, depending on whether the activity can be performed and how much difficulty is required. The endurance items are categorized into ordinal scale scores.

Recall period for items. Scoring of the CMAS requires that the activity be performed at the time the CMAS is administered (i.e., no recall period).

Endorsements. The CMAS has been included as a core set activity measure by both the International Myositis Assessment and Clinical Studies Group (IMACS) (1) and Paediatric Rheumatology International Trials Organisation (PRINTO) (2) for juvenile IIM. The CMAS (or alternatively, Manual Muscle Testing [MMT]) is also part of PRINTO's preliminary response criteria for juvenile dermatomyositis (DM) for the evaluation of muscle strength (4).

Examples of use. The CMAS has been used in validation and natural history studies (2,12,32–34) and is currently being used as a core set or ancillary outcome measure in several juvenile and adult polymyositis (PM)/DM therapeutic trials.

Practical Application

How to obtain. The CMAS is available from the original publication free of charge (31). It is also available from a variety of web resources, including the American College of Rheumatology web site (http://www.rheumatology.org/practice/clinical/pediatric_assessments/cmas.pdf) and the IMACS web site (http://www.niehs.nih.gov/research/resources/collab/imacs/diseaseactivity.cfm), along with detailed instructions and a training videotape.

Method of administration. The CMAS is administered by a trained therapist or clinician while observing the patient.

Scoring. Each item of the CMAS is scored depending on whether the activity can be performed and how much difficulty it requires. Scores of individual items range from 0–2 to 0–6, depending on the item. The CMAS can be easily scored by hand.

Score interpretation. The total CMAS score ranges from 0–52, with 52 representing normal or near-normal strength, function, or endurance. Age- and sex-related normal values for children ages 4–9 years have been published for 9 of the items, which document that younger children might not be able to reach a score of 52 (4). Validated cut points have not been established. However, as part of a consensus process, it was agreed that values <15 represented severe disease (32). In another publication, using a process that compared CMAS values to Childhood Health Assessment Questionnaire (C-HAQ) scores, values corresponding to no, mild, mild to moderate, and moderate impairment were 48, 45, 39, and 30, respectively (30).

Respondent burden. Assuming that all items can be attempted, the CMAS takes 15–20 minutes to complete. Some of the activities may be challenging for the weak child, and the overall assessment can be physically demanding for some.

Administrative burden. The ~15 minutes it takes to administer the CMAS may be a limitation in a busy clinic.

Scoring takes <1 minute and can be done by hand. Training in the administration of the CMAS is preferred. Proper equipment is needed to complete the entire test, including access to a step stool and chairs of appropriate height. Access to a watch with a second hand is needed, and a floor mat is helpful for the comfort of patients completing items performed on the floor.

Translations/adaptations. None available at present. The CMAS has been validated and studied in juvenile IIM patients. There have been no studies to date in adult myositis patients, although unpublished experience suggests the CMAS can also be used in adult myositis patients (Rider LG: unpublished observations).

Psychometric Information

Method of development. The 14 items of the CMAS were taken from and/or adapted from 2 unpublished clinical tools used by authors of the original CMAS publication (31). In this process, items from the 2 tools were reviewed by a group of pediatric rheumatologists, as well as a physical therapist and a physiatrist. Through consensus and observation of children with juvenile IIM attempting candidate items, the resulting 14-item tool was arrived at. Development of the scoring of each item was not described (31).

Acceptability. Although the tool is administered by the therapist or clinician, missing data can be common in children ages <5 years because of their limited ability to cooperate. Inability to complete a task is scored as 0. There are recognized ceiling effects (little change as children approach normal strength).

Reliability. *Internal consistency.* Not available.

Test–retest reliability. For juvenile IIM patients evaluated by trained assessors who evaluated the same patients in the morning and again in the afternoon, the Pearson's correlation coefficient for the total scores for each assessor pair ranged from 0.97–0.99 (all $P < 0.001$) and was 0.98 for the overall correlation of all assessors (31).

Interrater reliability. In juvenile IIM patients evaluated by 2 assessors, the intraclass correlation coefficient of the total score was 0.89 (very good) (30). In patients with juvenile IIM evaluated by 12 assessors, Kendall's W for each item ranged from 0.77–1.0 (all $P < 0.001$) and was 0.95 for the total score (31).

Validity. *Content validity.* This has not been formally assessed in juvenile IIM.

Construct validity. In children with juvenile IIM, the CMAS correlated highly with the C-HAQ and total MMT score (Spearman's $r = -0.73$ and 0.73, respectively, $P < 0.0001$) and moderately with physician global disease activity, physician skin activity, and parent disease severity, as well as serum creatine kinase and prednisone dose (Spearman's or Pearson's $r = -0.61$ to -0.44, $P < 0.0001$) (30,31). Correlations with magnetic resonance imaging of muscle edema and damage were moderate (Spearman's $r = -0.57$ to -0.48), and correlations with serum levels of enzymes were low but often significant (Spearman's $r = -0.36$ to -0.11). Correlations with the 10-cm visual analog scale (VAS) for physician global disease damage and physician skin disease damage were appropriately low (Spearman's $r = -0.15$ to -0.02, $P > 0.01$) (30).

Finally, in an international study of juvenile IIM, the CMAS correlated moderately with the Disease Activity Score (Spearman's $r = -0.54$), 10-cm VAS of parent overall disease severity (Spearman's $r = -0.56$), and physical summary score of the Childhood Health Questionnaire (Spearman's $r = 0.61$), and correlated highly with the C-HAQ (Spearman's $r = -0.71$) (2).

Criterion validity. Not available.

Ability to detect change. In children with juvenile IIM reassessed 7–9 months later, the overall standardized response mean (SRM) was 0.42 (95% confidence interval [95% CI] 0.21–0.63) (30). When those children with a 0.8-cm improvement in physician global disease activity were considered, the SRM was 0.89 (95% CI 0.53–1.09) (30). Finally, in children with juvenile IIM enrolled at diagnosis or requiring an escalation of therapy and reassessed 6 months later, the SRM was 1.4 (95% CI 1.2–1.5) (2).

A group of adult and pediatric rheumatologists and neurologists has reached consensus that measures of physical function should improve by ≥15% to classify a patient as improved for patients with juvenile and adult PM/DM (33). An absolute value for the minimum clinically important difference has not been determined.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The data demonstrate that the CMAS is a valid measure of strength, physical function, and endurance, which are of great importance to patients, families, and care providers. The requirement that the child being assessed is observed reduces the likelihood of biases in reporting. This instrument has excellent reliability, construct validity, and responsiveness in juvenile myositis for patients ages 4–18 years.

Caveats and cautions. Some clinicians believe that the CMAS takes too much time to administer. There are some concerns about ceiling effects. Appropriate training is necessary to reduce variability in assessments. The CMAS is difficult to assess in the youngest children with limited ability to cooperate. Like other measures of strength and function, the CMAS does not discriminate between activity and damage, and it may have poor sensitivity and specificity as a measure of activity for patients with moderate to severe damage, including patients who have muscle atrophy and fixed joint contractures. The CMAS has been validated and studied in juvenile IIM but not in other myositis subgroups.

Clinical usability. For some clinicians, the time required for administration limits the usefulness of the CMAS in the clinical context. However, others have found the CMAS extremely useful, particularly for longitudinal monitoring of patients.

Research usability. The CMAS is well suited to use in research. Concerns about ceiling effects may mean that it should be used with caution in patients with milder disease.

MYOSITIS DISEASE ACTIVITY ASSESSMENT TOOL (MDAAT)

Description

Purpose. The MDAAT is a tool that assesses disease activity of extramuscular organ systems and muscle to assess patients with adult and juvenile dermatomyositis (DM), polymyositis (PM), and inclusion body myositis (IBM). The MDAAT is a combined tool that includes the Myositis Disease Activity Assessment visual analog scale (VAS) (MYOACT) and the Myositis Intention to Treat Activities Index (MITAX). The MYOACT is a series of physician's assessments of disease activity in various organ systems using a VAS to assess the severity of activity that has been modified from the Vasculitis Activity Index (35), and the MDAAT is based on an intent-to-treat approach and modified from the British Isles Lupus Assessment Group (BILAG) approach to assess disease activity in lupus (36). The MITAX was published in 2004 (14) and updated in 2008, wherein items from the MITAX that were rarely scored were removed, glossary definitions clarified, and the criteria for scoring interstitial lung disease altered (37). The key issue in relation to the MDAAT is to ensure that the items recorded are, in the view of the physician, actually due to the active myositis and not due to disease damage, another unrelated disease process, or a side effect of medication.

Content. The MITAX assesses specific manifestations in 7 organs/systems, including constitutional, cutaneous, skeletal, gastrointestinal, pulmonary, cardiac, and muscle. The MYOACT consists of a 10-cm VAS for each organ system to score the overall severity of activity in each and a global extramuscular VAS.

Number of items. For the MYOACT, each organ system has a single VAS; a global extramuscular activity VAS is also scored. The VAS are anchored at the end points and midpoint. For the MITAX, 3–9 items consisting of symptoms/physical findings or laboratory abnormalities are assessed in each of the 7 organs/systems.

Response options/scale. For the MYOACT, the scores range from 0–10 cm. For the MITAX, each question is answered as 0 = not present, 1= improving, 2 = the same, 3 = worse, or 4 = new.

Recall period for items. Within 4 weeks.

Endorsements. Extramuscular activity has been considered by the International Myositis Assessment and Clinical Studies Group (IMACS) to be a core set activity domain, and the MDAAT is considered a validated tool to assess this domain in patients with adult and juvenile PM/DM (6). The MDAAT (either MYOACT or MITAX) is accepted by the Paediatric Rheumatology International Trials Organisation as a core set measure to assess the core set domain of global disease activity tool (2). The extramuscular activity from the MYOACT or MITAX is part of the preliminary criteria for response for adult and juvenile PM/DM (4,6).

Examples of use. The MDAAT has been used in natural history studies with the purpose of validating the tool (14,37), in studies examining disease activity (38), and as an outcome measure in therapeutic trials for adult and juvenile PM/DM.

Practical Application

How to obtain. The paper version is available at no cost. The tool is posted on the IMACS web site (http://www.niehs.nih.gov/research/resources/collab/imacs/diseaseactivity.cfm), along with slide sets for the cutaneous section and sample cases for training in scoring. The British Lupus integrated program (BLIPs) package, which includes activity measures for both lupus and myositis, can be obtained from Gordon Hamilton (e-mail: gordon.hamilton@limathon.com or Limathon@aol.com) (39). The cost of the computer version depends on the type of usage (commercial or academic). For further information about BLIPs, please contact Professor David Isenberg (e-mail: d.isenberg@ucl.ac.uk).

Method of administration. Clinician-completed, in-person administration based on history and examination.

Scoring. For the MYOACT, scores for each organ system and the extramuscular global activity are derived by measuring the distance the vertical line is from the left-hand side of the horizontal VAS. The length of the VAS should also be measured, so that the score can be adjusted to a denominator of 10 cm. For the MITAX, each clinical feature is recorded using a scale of 0–4, where 0 = not present, 1 = improving, 2 = the same, 3 = worse, and 4 = new. This score is then converted by a scoring schema to a final score ranging from A–E for each system, where A indicates very active disease requiring treatment with high-dose daily corticosteroids or a significant immunosuppressive therapy, B indicates a need for modest doses of corticosteroids and/or ongoing immunosuppression, C indicates a need for low-dose steroid or symptomatic drugs only, D indicates that the system is no longer active, and E indicates that the system was never active. Each organ system receives only a single A–E score (which can be numerically converted to A = 9, B = 3, C = 1, and D/E = 0 to obtain a global score) based on the score of the most severe item in that organ system. There has been work that has reassessed the scoring in lupus that may impact the scoring of the MITAX in the future (40). The tool can be scored by hand, but the BLIPs computer package can be obtained to convert the clinical assessments and provide the MITAX score.

Score interpretation. For the MYOACT, each organ system is scored from 0–10, and the 6 extramuscular organ systems can be summed to obtain an extramuscular score of 0–60, or a total score that includes the muscle system that ranges from 0–70. For the MITAX, the organ system scores are summed to obtain a total MITAX score with a range of 0–63, or 0–54 when the muscle system is excluded. The MITAX A–E organ system scores are intended to correspond with therapeutic choices for the patient, based on their level of disease activity. Normative data are not available.

Respondent burden. Not applicable.

Administrative burden. A complete history and physical examination are needed. To assess a patient in remission or close to remission takes <5 minutes. For a patient with a complex condition and who is not well known to the physician, it can take up to 15–20 minutes. For scoring, the BLIPs program can be run in the clinic or at a later time in ~5–7 minutes. Hand scoring may take a few extra minutes. Training using the resources on the IMACS web site is helpful.

Translations/adaptations. Only an English language version is available for both the paper and computer versions. The measure was developed and validated specifically for patients with inflammatory muscle disease, particularly for adult and juvenile PM/DM, although it should also be applicable to patients with IBM.

Psychometric Information

Method of development. The MITAX and MYOACT tools were developed from the BILAG for lupus and the Vasculitis Activity Index. Study evaluation forms from the Juvenile DM Disease Activity Study Group were used to develop the content of the subscales and some of the items, including adoption of elements of the Cutaneous Assessment Tool to the cutaneous organ system. The draft versions of the MITAX and MYOACT, including the glossary, were commented on and further refined by >75 members of IMACS using a Delphi approach. Two interrater reliability exercises using adult and juvenile PM/DM patients were performed that resulted in further refinement to the tool based on ease of use and understanding of the experienced adult and pediatric specialists who participated (14). During the course of a large multicenter study of adult PM/DM patients, the tool was further refined to improve the criterion validity (37).

Reliability. *Internal consistency.* In a natural history study of adult PM/DM patients to validate the MDAAT, correlation between the MYOACT and MITAX instruments for the individual organ systems was good, with correlation coefficients ranging from 0.80–0.94 (37).

Test–retest reliability. Not available.

Interrater reliability. In the initial study of adult PM/DM patients assessed by 7 raters, the reliability was considered good (with an intraclass correlation coefficient [ICC] of >0.65 or the ratio of the estimates of the standard error attributable to the physicians to the standard error attributable to the patients <0.40) for each of the organ systems of the MYOACT and MITAX, except for the MITAX constitutional system and the total MITAX score (14). Pediatric rheumatologists assessed juvenile DM patients, and the interrater reliability was also generally good, except for the skeletal system of the MITAX (6). The reliability studies were performed with prior training in the use of the tool and in the assessment and scoring of myositis activity.

Reliability was demonstrated in a 2-phase study of adult myositis patients evaluated in 7 centers and subsequently in patients reevaluated in 2 centers by 2 physicians at each center. The ICC was ≥0.6 in 5 of the 7 organ systems of the MYOACT and MITAX, as well as the total MITAX score, indicating generally good rater agreement. The mucocutaneous system of the MYOACT had the poorest interrater reliability (ICC 0.205) (37).

Validity. *Content validity.* Content validation is described above in Method of development.

Construct validity. From a large study of adult patients with PM/DM, the total MITAX score correlated moderately with physician global activity *(*Spearman's r = 0.69). The muscle MYOACT score also correlated moderately with the serum creatine kinase level (Spearman's r = 0.61) (37). In a separate study, the arthritis MYOACT and MITAX scores correlated moderately with Jo-1 autoantibody titers as a surrogate measure of disease activity (Spearman's r = 0.39–0.42), and mildly but significantly with the muscle MYOACT and MITAX scores, as well as with the total MITAX score (Spearman's r = 0.30–0.37) (38). In a study of juvenile idiopathic inflammatory myopathy patients and studies of treatment-refractory adult PM/DM patients, the MYOACT extramuscular global activity score correlated moderately with other core set measures of disease activity, including physician global activity, Manual Muscle Testing, Childhood Myositis Assessment Scale, and Childhood Health Assessment Questionnaire (Spearman's r = 0.29–0.54) (6).

Criterion validity. The criterion validity of the tool was measured by comparing the MITAX A score to the gold standard, defined as starting or increasing disease-modifying therapy in patients with adult PM/DM. The overall sensitivity and specificity in obtaining an A score on the MITAX index was 86% overall, with a specificity of 92%. The positive predictive value for a MITAX grade A score was 67% overall (37).

Ability to detect change. In a study of juvenile DM patients who were close to diagnosis or in need of disease-modifying therapy, the MYOACT extramuscular global activity score had a standardized response mean (SRM) of 1.3 (95% confidence interval [95% CI 1.1–1.5). The SRM for the total MITAX score was 1.2 (95% CI 1.0–1.3). In this same study, the MYOACT extramuscular global activity score and the total MITAX score showed good discriminant validity between patients who were rated as improved versus those who had not improved at 6-month reevaluation (2). In treatment-refractory adult PM/DM patients enrolled in therapeutic trials, the SRM was −0.4 but improved to −1.2 in patients who met the criteria for therapeutic response (6).

A group of adult and pediatric rheumatologists and neurologists has reached consensus that extramuscular activity should improve by ≥20% to classify a patient as improved in patients with juvenile and adult PM/DM (6). An absolute value for the minimum clinically important difference has not been determined.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The MDAAT, consisting of the MYOACT and the MITAX, provides the only in-depth disease activity score that captures a variety of organ systems that comprise extramuscular involvement. The muscle system as part of the full tool also comprises an integrated disease activity tool. Both the MITAX and MYOACT have excellent content validity, with a large amount of input in their development from myositis researchers and based on reliability study data. They also have good interrater reliability, moderate construct validity, and excellent responsiveness in adult and juvenile PM/DM patients (ages 2–18 years).

Caveats and cautions. The tool has been criticized by clinicians less experienced with myositis as being difficult to understand and score. However, in essence, the tool facilitates clinicians' asking their patients a comprehensive series of questions related to their disease, recording the symptoms as absent, better, same, or worse compared to the previous month. Training and experience with myositis patients clearly improve the reliability. Examination of previous MYOACT scores should reduce measurement error on serial evaluation. The VAS may be subjective and somewhat dependent on the experience of the rater. Although the MDAAT is recommended for use in patients with IBM, it has not been formally validated in this subgroup.

Clinical usability. The criterion validity of the MITAX A score supports use in the clinical setting. The time to administer the tool would not be much greater than a routine clinical assessment, but the burden is greater in complex patients or in patients with whom the physician lacks familiarity.

Research usability. The psychometric properties support its use in research studies and therapeutic trials. Training in the administration and scoring of the tool is important to improve reliability.

DISEASE ACTIVITY SCORE (DAS)

Description

Purpose. The DAS was developed to assess overall disease activity in juvenile dermatomyositis (DM) (41,42). The tool assesses muscle and cutaneous manifestations, including vasculopathic features, based on bedside clinical assessment.

Content. The DAS consists of 19 items, resulting in a score of 0–20: 10 items are scored dichotomously (the indicator is present or not) and 3 polychotomously (rating severity level or extent to which the indicator is present). In addition to the total score, it is also possible to report the DAS skin score (range 0–9) and the DAS muscle score (range 0–11) separately. According to the authors, the approximately equal contribution of items relating to muscle and skin reflects their equal importance in the disease pathophysiology.

Number of items, response options, and scoring. The presence or absence of weakness is assessed via 8 variables: neck flexor muscles, abdominal muscles, upper extremity proximal muscles, lower extremity proximal muscles, Gower's sign, abnormal gait, difficulty swallowing, and nasal speech. Functional status consists of a 4-point scale, ranging from normal function to severe limitations in daily life functions. The presence or absence of vasculitis is assessed by determining the presence of any 1 of the following: eyelid erythema, eyelid vessel dilation, eyelid thrombosis, nailfold erythema, nailbed telangiectasia, dilation of blood vessels on the palate, and "other" vasculitis. The presence of rashes is rated using polychotomous scales: the distribution of the involved skin is rated on a 4-point scale, ranging from none to generalized, while the severity of skin involvement is rated on a 5-point scale, ranging from absent to severe. Gottron's papules are rated on a 4-point scale, ranging from absent to severe, including evidence of atrophic lesions (which usually disappear entirely but can sometimes flare).

Recall period for items. The DAS refers to the status of the patient, as assessed by a trained health professional, on the day of the clinic visit. There is no recall period.

Endorsements. The DAS has been endorsed by the Paediatric Rheumatology International Trials Organisation (PRINTO) (8) as one of the 6 core set disease activity measures to be used to evaluate response to therapy in juvenile DM (2,4,43). Although the DAS has not been endorsed by the International Myositis Assessment and Clinical Studies Group (IMACS) as a core set measure, the group has recommended that it be included in future studies assessing outcomes and outcome measures for adult and juvenile myositis (http://www.niehs.nih.gov/research/resources/collab/imacs/main.cfm). The DAS, as a global disease activity tool, is also part of the PRINTO preliminary response criteria for juvenile DM (4).

Examples of use. The DAS has been used in validation and in natural history studies of juvenile DM, and has been incorporated as an end point in therapeutic studies (2,41–44).

Practical Application

How to obtain. The DAS is published and can be used free of charge for not-for-profit studies (42). The tool is also publicly available on the IMACS web site (http://www.niehs.nih.gov/research/resources/collab/imacs/main.cfm), along with instructions for administering the tool and training materials for the skin assessment.

Method of administration. Clinician-completed in-person administration.

Administrative burden. No information is available on the time to complete the questionnaire, but based on clinical use it takes 5–15 minutes to complete. The DAS can be completed by a physician or an allied health professional with adequate training.

Scoring. The total score ranges from 0–20, with the skin subscore ranging from 0–9 and the muscle subscore ranging from 0–11.

Score interpretation. A higher score indicates more active disease. Although normative data are not available, a normal score would be 0.

Translations/adaptations. None available at present. The DAS has been studied in patients with juvenile DM but not in other subgroups of patients with myositis.

Psychometric Information

Method of development. The DAS was developed at the Juvenile Myositis Clinic at Northwestern University Medical School's Children's Memorial Hospital (Chicago, Illinois) with the goal to rapidly assess how each child's clinical status has evolved over time (41).

Acceptability. The questionnaire is simple and easy to score. No specific information on the rate of missing data is available. However, in the PRINTO study, it was possible to calculate the DAS score in 99.3% of 275 patients (2).

Reliability. *Internal consistency.* The DAS produces a reliable estimate of disease activity (person separation = 2.80 compared with the criterion of 2.00) that distinguishes at least 3 distinct strata of disease activity in the sample: high, average, and low. The separate skin and weakness measures were less reliable, suggesting that both components are needed to adequately measure disease activity (41). Although the ratings across items were internally consistent, differences in practitioner sensitivity and specificity to individual disease activity indicators were found (42).

Interrater reliability. Using cutoffs of 0.40 and 0.20 to identify good and marginal agreement, respectively, 6 of the items for which coefficients could be estimated had good agreement, 6 had marginal agreement, and 4 had poor agreement as estimated by kappa coefficients (41). For most cases (~80%), the estimated disease activity measures were essentially the same across different physician raters. This result was confirmed by a Pearson's correlation coefficient of 0.79 between the 2 estimates of disease activity for each patient (41).

Test–retest reliability. Not available for juvenile DM.

Validity. *Content validity.* The fit of the DAS items to the disease activity construct is within acceptable levels (fit statistic values <1.30). Additionally, the relationship between measures of muscle strength and weakness is strong and negative (r = −0.77), with more strength (as rated by therapists using Manual Muscle Testing) being highly associated with less weakness (as rated by physicians using the DAS). The relationship between measures of disease activity and disability is weak (r = 0.20) (41).

Construct validity. The Spearman's correlation coefficients for the baseline to 6-month change in the DAS with the remaining 5 PRINTO/American College of Rheumatology/European League Against Rheumatism juvenile DM core set measures (physician's global activity assessment, Childhood Myositis Assessment Scale, Childhood Health Assessment Questionnaire, parent's global assessment of the patient's overall well-being, and Childhood Health Questionnaire physical summary score) were in the moderate range (Spearman's r = 0.4–0.6) (2). The DAS correlated moderately with other core set measures of disease activity (Spearman's r = 0.42–0.6) (2). The DAS skin score, but not muscle score, correlated weakly with periungual capillary loss (end row loops Spearman's r = −0.36) as well as with serum levels of muscle enzymes (42).

Criterion validity. There is no gold standard by which to establish criterion validity.

Ability to detect change. In the PRINTO study of juvenile DM, in a population requiring the initiation of new therapies, the standardized response mean of the DAS was 1.7 (95% confidence interval [95% CI] 1.5–1.9) (2). The DAS demonstrated significant ability to discriminate among patients who improved or did not improve at 6-month followup based on the physician's or parent's assessment of the child's response to therapy (2).

In the final logistic regression model of the PRINTO juvenile DM core set measures' ability to predict improvement, the physician's global assessment of the patient's overall disease activity and the DAS appeared to be the strongest predictors of response to therapy, with odds ratios of 3.4 (95% CI 1.5–7.4) and 3 (95% CI 1.4–6.5), respectively (2).

In a study of juvenile DM patients seen in followup, periungual nailfold capillary dropout was moderately associated with the skin DAS score ($\beta = -0.159$, $P < 0.0001$) and more modestly associated with the muscle DAS score ($\beta = -0.044$, $P < 0.0001$) (44).

The minimum clinically important difference has not been established for juvenile DM.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The DAS evaluates muscle weakness and skin disease activity, in particular both erythematous and vasculopathic rashes, in patients with juvenile DM. The DAS has been established as one of the 6 juvenile DM core set of measures of disease activity established by PRINTO, as it is a disease-specific global tool (8). The DAS was selected for use as a core set measure because of its superior responsiveness to clinically important change (and minor skewness) compared with the Myositis Disease Activity Assessment and Myositis Intention to Treat Activities Index; moreover, the DAS was the only index that used the entire range of possible scores (median score at baseline 12, range 0–20). The DAS has good internal consistency and construct validity and excellent responsiveness, but moderate to poor interrater reliability in patients with juvenile DM (ages 2–18 years).

Caveats and cautions. Several areas of the DAS are noteworthy for potential problems: the muscle weakness and function component, like all other measures of weakness and function in myositis, consists of a combination of both activity and damage indicators that may have poor sensitivity and specificity as an activity measure for patients with moderate to severe damage. Atrophic skin rashes are similarly scored, yet are considered a measure of damage rather than activity. The DAS does not capture involvement of all organ systems and has been studied in patients with juvenile DM, but not in other myositis subgroups.

Clinical usability. While the DAS is relatively simple to use with training and has overall good psychometric properties in patients with juvenile DM, the clinical meaning of scores has not yet been established, making this tool difficult to apply to the care of individual patients.

Research usability. The DAS has been well validated for juvenile DM, and given its psychometric properties and ease of use with training, it is appropriate for use in the research setting. The clinical meaning of DAS scores and clinically meaningful change in scores have yet to be established in the context of therapeutic trials (43). Studies of the DAS are needed in other myositis subgroups.

SHORT FORM 36 (SF-36)

Description

Purpose. The SF-36 is a widely used tool that assesses the global medical quality of life, functional health, and well-being of general and specific populations. The SF-36 is covered in detail in the article in this issue on adult measures of general health and health-related quality of life for further information. This section will cover only information on the SF-36 that is specific to myositis.

Endorsements. The SF-36 has been proposed by the International Myositis Assessment and Clinical Studies Group (IMACS) (http://www.niehs.nih.gov/research/resources/collab/imacs/abouttools.cfm) as an important patient-reported outcome measure to be used to evaluate response to therapy in all forms of myositis (1).

Examples of use. The SF-36 has been used in several small natural history studies of polymyositis (PM), dermatomyositis (DM), amyopathic DM, overlap myositis, and inclusion body myositis (IBM) (23,45–47), and in 2 exercise studies of PM and DM (48,49).

Practical Application

Translations/adaptations. The SF-36 is now available in many different languages (for details, e-mail info@iqola.org). It has been studied in a limited way in relatively small numbers of patients with adult PM, DM, amyopathic DM, overlap myositis, and IBM, but has been extensively studied in many other chronic diseases. The SF-36 is not recommended for use with children.

Psychometric Information

Reliability. Data are not available in myositis.

Validity. *Content validity.* None available in myositis.

Construct validity. In adult PM, DM, or overlap myositis, the physical functioning domain of the SF-36 correlated highly with the Health Assessment Questionnaire (HAQ) disability index (r = −0.71), whereas the HAQ correlated moderately with other domains of the SF-36, including role function, bodily pain, and emotional domain (r = −0.52 to −0.42). Manual Muscle Testing (MMT) scores, but not physician global activity, correlated moderately with SF-36 physical functioning, role functioning, and bodily pain (r = −0.57 to −0.27) (23). For patients with IBM, the physical functioning domain of the SF-36 correlated strongly with MMT, timed stand, timed walk, and the Amyotrophic Lateral Sclerosis Functional Rating Scale (46). In patients with adult DM or amyopathic DM, SF-36 subscales, including physical functioning, role functioning, physical, bodily pain, and general health, correlated mildly to moderately with physician global activity (Pearson's r = 0.30–0.42), and the social functioning, role functioning, and emotional and mental health domains of the SF-36 correlated more strongly with the Skindex emotion subscale (Pearson's r = 0.52–0.63). There was also a moderate negative correlation between grip force and the SF-36 health-related quality of life dimensions vitality and mental health in women with DM and PM (Spearman's r = −0.53 to 0.48) (47).

Criterion validity. In several studies from different countries, the SF-36 overall scores, and most or all of the 8 domain subscores, were significantly lower in patients with adult DM, PM, and IBM than in the general population. The physical functioning and role functioning domains were particularly impaired in myositis patients (23,45–47). Patients with chronic progressive illness had significantly greater bodily pain than those with relapsing–remitting illness (45).

Ability to detect change. Responsiveness statistics are not available in myositis.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SF-36 is a widely used and easily administered tool that is available in many languages. It has shown evidence of content, concurrent, criterion, construct, and predictive validity in many different chronic diseases, and extensive normative data are available. It is also recommended by IMACS as an important measure to assess patient-reported outcomes in all forms of adult idiopathic inflammatory myopathy. It has good construct and content validity in adult DM, PM, and IBM patients, and is not applicable to children with idiopathic inflammatory myopathy.

Caveats and cautions. The major drawbacks of the SF-36 are its limited use to date in myositis and the inconvenience and cost associated with obtaining a license to use it. Additional studies in all myositis subgroups are needed to more fully validate the tool and understand its role in assessing quality of life in myositis, particularly the reliability and responsiveness of the SF-36. The availability of recent variations of the SF-36, including the SF-36 version 2, SF-12, and SF-8, complicates the decision of which version to use in a given study.

Clinical and research usability. The SF-36 is easily administered to patients and is easily scored, making it appropriate for both clinical and research use. However, its cost may limit its use. The lack of data on responsiveness in myositis patients is a limitation for its use in myositis therapeutic trials.

CHILD HEALTH QUESTIONNAIRE (CHQ)

Description

Purpose. The CHQ, originally developed in the US in 1996, is a generic instrument administered to both parent and child designed to capture the physical, emotional, and social components of health status in children ages 5–18 years (50). As a generic questionnaire it can be used across different childhood conditions, and it has also been validated for use in juvenile dermatomyositis (DM) (28). The general content of the tool was discussed in other sections of this issue (see the article on measures of health status and quality of life in juvenile rheumatoid arthritis); therefore, only information specific to myositis will be discussed here.

Content. The CHQ consists of 14 health concepts: global health, physical functioning, role/social limitations–emotional/behavioral, role/social limitations–physical, bodily pain/discomfort, behavior, general behavior, mental health, self-esteem, general health perception, parent impact–emotional, parent impact–time, family activities, and family cohesion. In addition, there are 2 summary measures, the physical summary score (PhS) and the psychosocial summary score (PsS).

Endorsements. The CHQ PhS has been selected by the Paediatric Rheumatology International Trials Organisation (PRINTO; online at www.printo.it) (51) as a core set of measures to be used to evaluate response to therapy in juvenile DM (2,4,8). The CHQ, as an assessment of health-related quality of life (HRQOL), is also part of PRINTO's preliminary response criteria for juvenile DM (4). The International Myositis Assessment and Clinical Studies Group has proposed that HRQOL is an important patient-reported outcome measure to be used to evaluate response to therapy in all forms of myositis (1).

Examples of use. The CHQ has been used in validation and in natural history studies of juvenile DM (2,28).

Practical Application

Translations/adaptations. The CHQ is now available in 70 different languages (for details, see www.healthactchq.com), with 32 versions cross-culturally adapted and validated by PRINTO (29,52). The CHQ has been studied in patients with juvenile idiopathic arthritis, juvenile DM, and other chronic childhood diseases. Because it is a pediatric tool, the CHQ is not appropriate for use in adult myositis subgroups.

Psychometric Information

Method of development. Data regarding psychometric issues are extensively reported in the CHQ manual and can also be found in the supplement published by PRINTO (29,52) for each of the 32 validated translations. The psychometric properties of the CHQ have been established mainly for juvenile idiopathic arthritis and are discussed in this issue in the article on measures of health status and quality of life in juvenile rheumatoid arthritis. However, data were further confirmed in a study that investigated the change over time of HRQOL in patients with active juvenile DM, as measured by the CHQ.

To appropriately evaluate the underlying framework and psychometric properties of the CHQ, PRINTO used item-scaling multitrait analysis software. Since the main validation analysis was conducted when the original English versions of the CHQ (28) were developed in the US, the PRINTO revalidation of the questionnaire was set up as "confirmatory," meaning that the PRINTO results were considered successful if they were equal to or superior to the results published for the original American English version of the CHQ.

Acceptability and reliability. This has not been assessed in juvenile DM.

Validity. *Content validity.* In a study by PRINTO, the mean ± SD CHQ PhS and PsS were significantly lower in juvenile DM patients than in healthy children (33.7 ± 11.7 versus 54.6 ± 4.1 and 45.1 ± 9.0 versus 52 ± 7.2, respectively), with physical well-being domains being the most impaired. In addition, both the PhS and PsS decreased with increasing level of disease activity and muscle strength, and inversely correlated with the parent's evaluation of the child's overall well-being. The study also showed that a Childhood Health Assessment Questionnaire (C-HAQ) score >1.6 (odds ratio [OR] 5.06), a child's overall well-being score >6.2 (OR 5.24), and to a lesser extent muscle strength and alanine aminotransferase level were the strongest determinants of poorer physical well-being at baseline, whereas baseline disability and longer disease duration were the major determinants for poor physical well-being at followup (28).

Construct validity. In terms of content validity, the CHQ correlates strongly with the C-HAQ (Spearman's r = −0.73) and moderately with the Childhood Myositis Assessment Scale (Spearman's r = 0.61) and other core set measures of disease activity (Spearman's r = −0.42 and −0.58 with physician and parent global activity and Disease Activity Score, respectively) in juvenile DM (2).

Criterion validity. There is no gold standard by which to establish criterion validity.

Ability to detect change. Responsiveness was tested specifically in juvenile DM, in which patients with active disease who needed to increase therapy were assessed at baseline and after 6 months. The standardized response mean of the PhS of the CHQ in this PRINTO study was 1.0 (95% confidence interval [95% CI] 0.9–1.2), whereas that of the CHQ PsS was 0.5 (95% CI 0.3–0.6) (2). The PhS of the CHQ did not have significant discriminant validity to separate juvenile DM patients whose disease was considered to be improved after initiation of new therapy from those whose disease did not improve (2).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. One of the 6 components of the juvenile DM core set established by the American College of Rheumatology/European League Against Rheumatism/PRINTO is the evaluation of the domain HRQOL, and the PhS of the CHQ has been suggested as a possible tool for evaluating that domain (other tools might also be used).

The CHQ has good content and construct validity and responsiveness in large studies of juvenile DM for children ages 2–18 years.

Caveats and cautions. The major limitations of the CHQ are its length and the fact that the parent version is mainly used for clinical research because the child version is too long to be used in research or clinical settings. Several other HRQOL scales are available for use in children with pediatric rheumatic diseases (53,54); however, most of them have remained essentially research tools and are not routinely administered in most pediatric rheumatology centers. There is a degree of redundancy between the PhS of the CHQ and the C-HAQ, as both are measures of physical function, although the CHQ has a broader construct in assessing HRQOL more generally (2).

Clinical usability. The psychometric evaluation would support interpretation of scores to make decisions for individual patients. Two reasons that this instrument is not commonly incorporated in standard clinical care are its length and complexity.

Research usability. The psychometric evaluation supports use of the CHQ for research studies of juvenile DM. The administrative and respondent burden may limit its use. Studies for other myositis subgroups are needed.

PHYSICIAN GLOBAL DAMAGE

Description

Purpose. An overall rating of the disease damage related to myositis, defined as persistent changes in anatomy, pathology, physiology, or function, such as fibrosis, scarring, or atrophy, resulting from any cause (including prior treatment) since the onset of the myositis. Features of damage, or the pathology that led to the feature, must be present for at least 6 months despite immunosuppressive or other therapy, including exercise and rehabilitation (1).

Content. The global assessment of disease damage is to be judged by the physician based on all of the information available at the time of the evaluation, including the subject's appearance, medical history, physical examination, laboratory testing, and the prescribed medical therapy. The global disease damage assessment is completed on a 10-cm visual analog scale (VAS) that is often anchored at the end points and middle.

Number of items. 1 item, either a VAS or a Likert scale rating.

Response options/scale. For the VAS rating, a score of 0–10 (down to 1 decimal place) is used, and for the Likert scale, a Medical Research Council grade of 0 (no disease damage), 1 (mild disease damage), 2 (moderate disease damage), 3 (severe disease damage), or 4 (extremely severe disease damage) is used. The 10-cm VAS may have better precision, sensitivity, and specificity, but the 2 scales highly correlate (5).

Recall period for items. The global disease damage score is based on a current assessment, although a recall period of up to 2–4 weeks for the components of global disease damage is acceptable.

Endorsements. The physician global disease damage has been recommended to be included in the assessment of damage for adult and juvenile patients with polymyositis (PM), dermatomyositis (DM), and inclusion body myositis (IBM) by the International Myositis Assessment and Clinical Studies Group (IMACS) (1), and achieved consensus to be included as a core set measure of disease damage for patients with juvenile DM by the Paediatric Rheumatology International Trials Organisation (8).

Examples of use. Natural history studies, particularly those validating the Myositis Damage Index (MDI) and other damage assessments (27,34), as well as several myositis therapeutic trials that have recently completed enrollment.

Practical Application

How to obtain. The physician global damage assessment is available in publications using this as an assessment tool, free of charge (5). The IMACS web site (http://www.niehs.nih.gov/research/resources/collab/imacs/diseasedamage.cfm) also hosts copies of these tools, including the grading scales and detailed instructions, along with example cases and sample ratings as training materials for physicians.

Method of administration. The physician global damage assessment is completed by the physician assessing the patient and includes factors involving the subject's appearance, medical history, physical examination, laboratory testing, and the prescribed medical therapy.

Scoring. A single score is derived by measuring the distance of the vertical line from the left end of the horizontal VAS. The length of the VAS should also be measured so that the score can be adjusted to a denominator of 10 cm. The Likert scale also results in a single score. Scoring takes <1 minute and is done by hand.

Score interpretation. 0 represents inactive disease, and the higher the score the more severe the disease damage.

Respondent burden. Not applicable.

Administrative burden. The time to complete the physician global damage assessment is <1 minute, but this requires integration with other assessment measures to derive an overall impression.

Translations/adaptations. The physician global damage assessment has been used internationally in the native languages of the patient and examiner (8,34). Physician global damage has been studied and used in adult and juvenile PM/DM, as well as a number of systemic rheumatic diseases.

Psychometric Information

Method of development. Physician global damage assessment was first used in the assessment of other systemic rheumatic diseases, including systemic lupus erythematosus and systemic vasculitis. It was then adopted and studied in myositis.

Acceptability. Missing data are not common, and floor and ceiling effects are not common. There can be measurement error if physicians do not look at their previous ratings as part of the determination of the current rating. Although based on the collection of objective data, the rating itself is subjective and based on the experience of the rater.

Reliability. *Internal consistency.* Regarding internal reliability, Spearman's correlation was excellent (Spearman's r = 0.89) for the correlation of the VAS to the Likert scale for physician global disease damage, and the intraclass correlation coefficient was 0.85 ($P < 0.0001$) (5).

Test–retest reliability. Not available.

Interrater reliability. In a study of pediatric rheumatologists assessing paper cases of juvenile DM, the κ coefficient for agreement with the Likert scale ratings of global disease damage was 0.76 and Cronbach's α was 0.98 (5).

Validity. *Content validity.* In validating the physician global activity, pediatric rheumatologists reached consensus that 4 variables (calcinosis, muscle atrophy, functional assessment, and joint contractures) were extremely important in the determination of juvenile DM global disease damage and that 16 clinical parameters were unimportant or mildly important in the assessment of damage (5).

Construct validity. In a natural history study of juvenile PM/DM patients, the physician global damage assessment strongly correlated with the total extent and severity of damage in the MDI (Spearman's r = 0.79–0.88) (27). In the same study, which also examined treatment-refractory adult PM/DM patients, the physician global damage assessment moderately correlated with the total extent and severity of damage in the MDI (Spearman's r = 0.42–0.82) (27).

Criterion validity. There is no gold standard upon which to assess criterion validity. Sometimes the physician global damage is used to assess criterion validity in studies validating other measures of damage.

Ability to detect change. In the juvenile idiopathic inflammatory myopathy natural history study of patients who were reassessed 8 months after study entry, the standardized response mean for physician global damage was poor at 0.02 for the Likert scale and 0.14 for the VAS scale (5).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The data demonstrate that physician global damage is a reliable measure of damage, with some content and construct validity in juvenile (ages 2–18 years) and adult PM/DM patients, and as expected, it has little responsiveness over a relatively short period of time (8 months).

Caveats and cautions. To reduce variability, this measure requires training of the person performing the assessment. The VAS may be subjective and somewhat dependent on the experience of the rater. Physician global damage has not been formally validated in IBM, and the validation data in PM/DM are limited.

Clinical usability. The measure should be useful in the assessment of myositis patients, particularly for longitudinal evaluation of patients over several years. Examination of previous measurements in formulating serial ratings should help reduce measurement error.

Research usability. Physician global assessment of damage is well suited to use in research and is becoming widely used in myositis long-term outcome studies and therapeutic trials. It is considered a core assessment of disease damage.

MYOSITIS DAMAGE INDEX (MDI)

Description

Purpose. The MDI scores damage, which is defined as persistent or permanent change in anatomy, physiology, and function that develops from previously active disease, complications of therapy, or other events (1). The MDI is patterned after the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index (SDI) (55,56) and is intended to be used in patients with adult and juvenile dermatomyositis (DM), polymyositis (PM), and inclusion body myositis (IBM).

Content. The MDI measures specific manifestations in 11 organ systems. The MDI also includes a series of visual analog scales (VAS) to quantify damage severity in a given organ system. The MDI is structured for both pediatric and adult patients, and certain items are scored solely in each population.

Number of items. There are 11 separate VAS ratings that constitute the MDI severity of damage scale. Individual items are assessed by the MDI extent of damage scale. There are 35 items in children, 37 in adolescents, and 38 in adults. There are also 16 optional items that require additional testing, which constitute the MDI extended damage scale.

Response options/scale. The 10-cm VAS are anchored at the end points and the midpoint. Each of the 11 organ systems has 3–6 items scored as present or absent.

Recall period for items. To receive a positive score, each item must be present for at least 6 months (or the pathology that led to the feature must have been present for at least 6 months) despite prior immunosuppressive or other therapy. Only items present since the date of diagnosis are included.

Endorsements. The MDI was developed by the International Myositis Assessment and Clinical Studies Group (IMACS) and is endorsed by IMACS to measure damage as an important outcome to be assessed in myositis research studies and therapeutic trials (1). The Paediatric Rheumatology International Trials Organisation has included the MDI as part of the preliminary core set of disease damage measures for the assessment of juvenile DM (8).

Examples of use. The MDI has been used in validation studies (14,27,57), as well as in long-term outcome studies (12,34,58,59).

Practical Application

How to obtain. The MDI is available on the IMACS web site (http://www.niehs.nih.gov/research/resources/collab/imacs/diseasedamage.cfm) and as part of the original publication (14). There is no cost associated with use of the paper version. The questionnaire is also available as part of the British Lupus integrated program software (39). The computer version is available from Gordon Hamilton (e-mail: gordon.hamilton@limathon.com or Limathon@aol.com), with an associated cost for commercial use.

Method of administration. Clinician-completed, in-person administration.

Scoring. For the VAS, scores for each organ system are determined by measuring the distance of the vertical line from the left-hand side of the horizontal VAS. The length of the VAS should also be measured so that the score can be adjusted to a denominator of 10 cm. For items in the damage index, the score is 1 point if present and 0 if absent. In order for an item to be scored as a damage item, the problem must have been present for at least 6 months and must be expected to persist or be irreversible and not treatable with immunosuppressive medication.

Score interpretation. The VAS are summed together for a potential score of 0–110 for the MDI severity of damage score. For each organ system, 0 = no damage and 10 = extremely severe damage. For the individual items, these are summed together to comprise the extent of damage score, with ranges of 0–35 in children, 0–37 in adolescents, and 0–38 in adults. The optional items comprise the MDI extended damage score, and these are summed together for a potential score range of 0–16. Missing items are scored as not assessed. The clinical meaning of MDI scores has not been established.

Respondent burden. Not applicable.

Administrative burden. A complete history and physical examination are needed. The rate-limiting factor is the accessibility to previous notes (paper or electronically obtained). To complete the form for a patient who is essentially well, scoring will take <1 minute. For a complex patient not known to the physician, it may take 20–30 minutes. Some training in the use of the tool is advisable. The IMACS web site provides some training materials, with sample cases and ratings, as well as a slide collection for the cutaneous manifestations of damage.

Translations/adaptations. The MDI is available only in English. The MDI has been used in patients with adult and juvenile PM/DM.

Psychometric Information

Method of development. The MDI was modified from the SDI (55,56). A 10-cm VAS for each organ system was also included to measure severity of damage. The draft version of the MDI, including the glossary, was commented on and further refined by >75 members of IMACS using a Delphi approach. Two interrater reliability exercises using adult and juvenile PM/DM patients were performed that resulted in further refinement of the tool based on feedback in ease of use and understanding of the experienced adult and pediatric specialists who participated (14).

Acceptability. Missing data are common in the MDI extended damage score, and that portion of the tool has not been formally validated. There are no known floor or ceiling effects, and in fact, most patients with adult and juvenile PM/DM have measurable damage several years after diagnosis (12,27,34,57–59).

Reliability. *Internal consistency.* In studies of juvenile and adult PM/DM, total MDI extent and severity of damage scores were highly correlated (Spearman's r = 0.87 in juvenile and 0.75–1.0 in adult PM/DM) (27,57).

Intrarater reliability. Not available.

Interrater reliability. In a study of adult patients with PM/DM, the reliability was considered good (with an intraclass correlation coefficient [ICC] of >0.65 or the ratio of the estimates of the standard error attributable to the physicians to the standard error attributable to the patients <0.40) for each organ system of the MDI extent and severity scores, except for the gastrointestinal and pulmonary systems for extent of damage and the skeletal system for severity of damage (14). Good interrater reliability for most organ systems was confirmed in a subsequent multicenter study of adult PM/DM, where the ICC values for the MDI severity and extent of damage scores ranged from 0.65–0.84, except for the gastrointestinal, cardiac and peripheral vascular, and malignancy systems, where the ICCs ranged from 0.20–0.56 (12).

Validity. *Content validity.* Content validation is described above in Method of development.

Construct validity. In a study of juvenile and adult patients with PM/DM, total MDI extent and severity of damage scores highly correlated with physician global damage (Spearman's r = 0.79–0.88). In juvenile patients with PM/DM, MDI severity of damage, as well as the muscle and skeletal system scores, also correlated moderately with the Childhood Health Assessment Questionnaire as a functional disability measure, with Manual Muscle Testing as a measure of strength, with the T1-weighted magnetic resonance imaging (MRI) score, and inversely with serum creatinine (Spearman's r = 0.37–0.58). These findings were replicated in additional studies of juvenile DM (12,58). In adult patients with PM/DM, only serum creatinine and T1-weighted MRI correlated with the muscle system severity of damage score (12). In adult PM/DM patients, there was moderate correlation of most organ systems between the MDI and the Myositis Intention to Treat Activities Index (Spearman's r = 0.33–0.73 for muscle, cutaneous, gastrointestinal, and pulmonary systems) and lower correlation in cardiac and skeletal systems (Spearman's r = 0.13–0.24) (57).

Criterion validity. In patients with adult or juvenile PM/DM, those with a chronic illness course had a higher rate of damage accumulation than those with a monocyclic or polycyclic course, and the percentage of patients with measurable damage was also greater in those with a chronic illness course (27). This finding was replicated in a large international study of juvenile DM (34). Adult patients with PM/DM who died had higher damage scores at last followup, including in the cardiovascular and pulmonary systems, than patients who remained alive (27).

Ability to detect change. In adult patients with PM/DM who had treatment-refractory disease, there was a measurable increase in the annual change in the total MDI severity of damage score, with a median increase of 2.4 points (whereas the annual rate of change in the total MDI extent of damage score was undetectable, median 0) (27). Patients with juvenile PM/DM, at a median of 80 months from diagnosis, had no detectable annual rate of increase in their damage scores (27). In juvenile DM patients close to the time of diagnosis, the mean increase in the MDI extent of damage score was 0.01 per 6 months in the 6 months after diagnosis (58). In 1 cohort of juvenile DM patients, MDI extent of damage scores improved in 65% of patients at last followup (59).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The MDI offers a comprehensive assessment of the potential consequences of having myositis, complications of treatment associated with myositis, and other potential contributions to morbidity. The MDI is constructed to measure both severity and extent of damage. From the preliminary validation studies, the severity of damage score might be more sensitive in detecting damage and more sensitive to change. Although the 2 scores correlate highly, it is recommended that both measures be used simultaneously. The MDI has good reliability, good construct validity, and excellent criterion validity in juvenile (ages 3–18 years) and adult PM/DM.

Caveats and cautions. The MDI does not measure only damage related to disease, but it also captures other comorbid conditions. Although damage scores are meant to reflect irreversible changes, improvement in some damage elements has been reported in children with juvenile DM. It is unclear whether the presence of an element for 6 months is long enough for it to represent damage or whether it might still be part of active disease, especially early in the course of illness. Training in the use of the tool and experience with myositis patients clearly improve the reliability. Examination of previous severity of damage VAS scores should reduce measurement error on serial evaluation. The VAS may be subjective and somewhat dependent on the experience of the rater. Although the MDI is recommended for use in patients with IBM, it has not been formally validated in this subgroup.

Clinical usability. The MDI may be useful to track damage and affected organ systems over time, but the scores have no determined clinical meaning.

Research usability. The MDI may be used in long-term observational studies or in clinical trials, mainly to see that patients treated with a new immunosuppressive therapy do not have increased damage over time. Certain novel therapies may be directed toward specific treatment of damage elements (such as treatment of calcinosis or muscle regenerative therapies), in which case the MDI can be an important outcome measure for such trials.

QUANTITATIVE MUSCLE TESTING (QMT)

Description

Purpose. To measure the amount of maximum isometric force generated from a muscle group using specialized equipment.

Content. In inclusion body myositis (IBM) studies, the following muscle strength measurements are typically tested: bilateral elbow flexion and extension, bilateral knee flexion and extension, bilateral ankle dorsiflexion, and bilateral grip strength.

Number of items. This ranges from 6 muscle groups tested bilaterally to 20 muscle groups tested bilaterally (creating 12–40 individual items). The individual muscle group results can be averaged across all muscle groups tested to create a composite score, which can then be converted to a Z score.

Response options/scale. In kilograms, with a range of up to 100 kg for each muscle group tested. Response is based on the strength of the muscle group being tested and the maximum load allowable on the tensiometer (100 kg).

Recall period for items. None.

Endorsements. None.

Examples of use. There have been several phase II trials of interferon-β for IBM (10,60), an ongoing phase II trial of arimoclomol in IBM (http://clinicaltrials.gov/ct2/show/NCT00769860), an etanercept in dermatomyositis (DM) trial (http://clinicaltrials.gov/ct/show/NCT00282880), an etanercept trial in IBM (61), an oxandrolone trial in IBM (62), intravenous immune globulin trials for IBM (63,64), and an alemtuzumab trial in IBM (65). Rose et al (66) conducted a prospective natural history trial that showed a 4% mean decline in composite strength score from baseline over 6 months. There are no validation studies of QMT in IBM, and a single validation study of handheld pull gauge to measure isometric dynamometry in polymyositis (PM)/DM patients (67).

Practical Application

How to obtain. QMT equipment for fixed-strength measurement can be purchased online at www.aeverl.com. The fixed device contains a tensiometer that the subject pulls against. The tensiometer is connected to a Zimmer frame device attached to an adjustable bed.

Method of administration. The position of the patient depends on the muscle group being tested. A strap is placed distal to the movement being tested. This strap is connected to the tensiometer, which is attached to a fixed location (i.e., Zimmer frame). There is tension in the strap and the tensiometer. The joint tested is placed in midrange position. The patient is asked to pull as hard as they can. There should not be any movement in the joint being tested (isometric force). For instance, for knee flexion and extension, the subject is sitting, and the knee is in 90° of flexion, with the strap at the ankle, above the lateral malleolus. If testing flexion, the strap is hooked to the tensiometer so that the patient can attempt to bend the knee. The patient has to be stabilized. A handheld pull gauge device is also available (67).

Scoring. Results range from 0–100 kg for each muscle group tested. A patient's log(QMT score) for a particular muscle group is standardized by subtracting his or her predicted score in the appropriate model, given the patient's age, sex, and height, and dividing by the SD around the fitted model (68). The resulting measurement can be interpreted as the number of SDs from average normal strength, after accounting for age, sex, and height. A composite QMT score for a patient is formed by averaging the standardized QMT scores across all muscle groups tested (69).

Score interpretation. Normative data have been obtained by recruiting from hospital personnel and family members as well as family members of amyotrophic lateral sclerosis (ALS) patients. The standardization process involved constructing regression models for the relationship between log(QMT score) and age, sex, and height among normal subjects for each muscle group separately (69). Normative data using other equipment systems are also available (70,71).

Respondent burden. Depends on the strength, fatigability, and effort of the patient.

Administrative burden. Up to 1 hour to test multiple muscle groups. Testing 1 or 2 muscle groups using a handheld device can take 15 minutes.

Translations/adaptations. Has been widely used in patients with IBM, with limited reliability data in adult DM and PM. It has also been widely used in other muscle diseases, such as muscular dystrophies and ALS.

Psychometric Information

Method of development. It was derived from studies of muscle strength deterioration over time in ALS.

Acceptability. Missing data are common. If a muscle group is missed due to an injury, the missing data are imputed in an intent-to-treat analysis that averages the values from the visit before and after the missing time point. The floor effect can be present in weaker patients: are they able to actively position the joint in the position to be tested or maintain that position until the test is completed? The ceiling effect is determined by the amount of strength the tensiometer can withstand.

Reliability. *Internal consistency.* There have been no internal consistency studies conducted in ALS or patients with myositis.

Test–retest reliability. In ALS, the intrarater test–retest correlation was 0.96 for normal controls and 0.98 for ALS patients. The mean absolute percent variation of testing and retesting was 6.5% for normal subjects and 8.9% for ALS patients (70). In Duchenne's muscular dystrophy (DMD), intrarater test–retest correlations ranged from 0.88–0.99 for children with DMD and from 0.85–0.98 for children without DMD (72). Interrater reliability ranged from 0.81–0.98 by analysis of variance (ANOVA) in a study of 13 muscle groups tested by a handheld pull gauge in patients with stable PM/DM (67). No studies have tested the reliability of QMT in patients with IBM.

Interrater reliability. The mean interrater test–retest correlation was 0.95 for normal controls and 0.98 for ALS patients. The absolute mean percent variation between QMT trials is 7.6% for healthy subjects and 8.2% for ALS patients (73). Interrater test–retest correlations ranged from 0.74–0.97 in children with DMD and from 0.71–0.98 in children without DMD (73). These numbers are similar in subjects with facioscapulohumeral dystrophy (FSH) (69). Intrarater reliability ranged from 0.88–0.98 by ANOVA in a study of 13 muscle groups tested by a handheld pull gauge in patients with stable PM/DM (67). No studies have tested the reliability of QMT in patients with IBM.

Validity. *Content validity.* No studies have been done to show validity in patients with PM, DM, or IBM.

Construct validity. In FSH, the correlation between the composite QMT score and Manual Muscle Testing (MMT) scores was strong (r = 0.88) (69). QMT was shown to correlate strongly with the Inclusion Body Myositis Functional Rating Scale (Pearson's correlation coefficient at baseline = 0.73 and then at 24 weeks = 0.80) (74).

Criterion validity. There is no criterion validity available in PM, DM, or IBM patients.

Ability to detect change. In FSH, both the QMT (P = 0.04) and MMT (P = 0.05) were able to detect changes in strength over time (69). Rose et al (66) demonstrated in a natural history study that the mean ± SD decline in composite strength score from baseline was 4% ± 5.8% over 6 months (P = 0.05), but that the rate of progression was variable and that 4 of the 11 subjects involved did not show any decline. Dalakas et al (75) reported a 14.9% in decline in strength in the Alemtuzumab (CAMPATH 1-H) study. The standardized response mean is not available for patients with myositis.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Isometric dynamometry provides a quantitative measure that might be sensitive in detecting small changes in strength as well as mild weakness that might not be detected by MMT.

Caveats and cautions. The person administering the test must be trained. There are many different instruments (hardware and software) to measure quantitative muscle strength. Some require more training than others. Also, depending on the unit, it may need a dedicated room to house the equipment. QMT is difficult to use on patients who have trouble moving or have less than antigravity strength. The cost of equipment can run to ~$15,000. Like other measures of strength, QMT does not discriminate between activity and damage and may diminish in sensitivity and specificity as an activity measure for patients who are farther along in their illness course with accumulated damage and progressive muscle atrophy. There is almost no validation in patients with myositis, including limited reliability data in adult PM/DM and limited construct validity in IBM patients. There are no data using QMT in juvenile idiopathic inflammatory myopathy patients.

Clinical usability. It takes ~1 hour to test a full set of muscle groups; therefore, it is not a good tool to use during routine clinic visits.

Research usability. QMT is difficult to use due to cost, training, and retraining of study personnel. QMT has been successful as an end point in IBM trials in detecting significant drug effects (62,75).

MYOSITIS FUNCTIONAL INDEX-2 (FI-2)

Description

Purpose. The FI-2 was developed as a disease-specific observational tool for adult patients with polymyositis (PM)/dermatomyositis (DM) to measure muscle endurance (76). The FI-2 is a more developed version of the original Functional Index (FI), which was presented in 1996 as the first disease-specific muscle impairment measure for patients with PM/DM (77).

Content. The FI-2 measures the number of repetitions performed in 7 muscle groups: shoulder flexion, shoulder abduction, neck flexion, hip flexion, and knee extension (step test; performed at a pace of 40 beats per minute, which is monitored by a digital metronome), and heel lifts and toe lifts (performed at a pace of 80 beats per minute). Each muscle group is scored as the number of correctly performed repetitions, with no total score, presenting a profile of muscle impairment for the upper and lower extremities and the neck.

Number of items. If the assessment is performed on both the right and left extremities, the FI-2 consists of 11 items, and when it is performed on the dominant body side, there are 7 items. There is no total score.

Response options/scale. Each muscle group is scored by the number of repetitions performed, and the score ranges from 0–60 for the shoulder flexion, shoulder abduction, neck flexion, hip flexion, and step test tasks, or from 0–120 for the heel and toe lifts.

Recall period for items. The patient performs the test and is observed and scored by a trained health professional. There is no recall period.

Endorsements. None.

Examples of use. The FI-2 is used in clinical practice in Sweden to measure muscle endurance of adult PM/DM patients at yearly followup visits and to assess changes after interventions such as exercise or medical treatment. The FI-2 has been used in 1 study evaluating a 7-week intensive resistance training program for patients with chronic PM/DM (78).

Practical Application

How to obtain. The protocol of the FI-2 and written instructions can be obtained at no cost in the original publication (76). The tool, as well as an instructional slide set and video, can be found on the International Myositis Assessment and Clinical Studies Group web site (http://www.niehs.nih.gov/research/resources/collab/imacs/main.cfm).

Method of administration. The FI-2 is a direct observational assessment tool.

Scoring. The number of correctly performed repetitions is recorded, together with the perceived muscle exertion for each task. Computer scoring is not necessary.

Score interpretation. The number of correctly performed repetitions is scored for each muscle group; they vary from 0–60 or 0–120, where 0 = severe limitation and 60 (or 120) = no limitation. After each muscle group is tested, the patient rates his/her perceived muscle exertion according to the Borg CR-10 scale, which ranges from 0–10, where 0 = no exertion and 10 = extremely strong, almost maximal exertion (79). The Borg CR-10 scale is not included in the FI-2 but is used to measure how much effort the patient exerts to complete each task. This enhances the observer's ability to detect whether the patient stops due to reasons other than muscle fatigue, such as pain or lack of motivation. To date, normative data are not available for the FI-2.

Respondent burden. Not applicable.

Administrative burden. The maximal time required to perform each muscle group is 3 minutes, and the maximal time to perform the FI-2 on both the right and left sides is 33 minutes. If the FI-2 is performed only on the dominant side, the time required is 21 minutes. The ratings of perceived exertion (the Borg CR-10) add an additional minute. The FI-2 takes ~5 minutes to score, and no training is required for scoring. In some centers, the FI-2 is performed in a separate session by a trained physical therapist.

Translations/adaptations. No translations or cultural adaptations are currently available. The tool has not yet been tested in other populations, only in adult PM/DM, with unpublished clinical observations in patients with inclusion body myositis (IBM).

Psychometric Information

Method of development. The FI-2 was based on the previous version, the FI, which was also developed specifically for patients with PM/DM (77). The FI assessed the number of repetitions (maximal number of repetitions, range 10–20) in elbow flexion, shoulder flexion, shoulder abduction, neck flexion, trunk flexion (sit-up), hip flexion, knee extension (step test) as well as heel lifts, and toe lifts performed standing on 1 leg at a time. The FI also included tests of grip strength using the Grippit instrument (80), ability to transfer from side to side and up to a sitting position, as well as peak expiratory flow. Patients and health professionals were involved in the validation process. Due to ceiling effects and problems with internal consistency with several items in the FI (discussed below), a group of health professionals and patients agreed to remove hip abduction, transfers, and peak expiratory flow from the tool when the FI-2 was created (76). Despite ceiling effects, the neck flexion and sit-up tasks were considered relevant. All tasks of the FI were functional tasks except for the grip strength, so the Grippit assessment was excluded from the FI-2, but it is recommended that it be assessed as a separate measure. The number of repetitions was increased to 60 or 120 for each task, and the dorsal and plantar flexion tasks were revised to be performed standing on both feet instead of balancing on 1 foot. To further ensure stability to the tasks, repetitions are performed at a specific pace guided by a metronome.

Acceptability. For the FI (version 1), ceiling effects, defined as the median value equaling the maximal score for each muscle group, were evident for 8 of the 11 muscle groups, the transfers, and the peak expiratory flow (76). No floor or ceiling effects have been found in patients with PM/DM with the FI-2, and the mean number of repetitions for each item varies from 60–120 in patients with PM/DM (76). However, clinical practice indicates that there might be floor effects when used in patients with IBM, especially the knee extension, the heel lift, and the toe lift tasks. There are generally no missing data with the tool, and if the patient will not attempt a particular item, the score is 0 on that item.

Reliability. *Internal consistency.* Because each muscle group is scored individually and not included in a subscale, internal consistency analysis is not relevant for the FI-2.

Test–retest stability. The measurement error for each task varies between 5 and 16% (76).

Rater reliability. The FI-2 demonstrated good to excellent intrarater reliability for all tasks, with intraclass correlation coefficients (ICCs) for the 7 tasks varying between 0.75 and 0.99 (76). Systematic variations were revealed for the shoulder flexion task, indicating that a training session for the patient is necessary to ensure good intrarater reliability. Interrater reliability was also good to excellent, with ICC coefficients of 0.86–0.99 for the tasks without systematic variation. It is advised that the assessor train on how to score the tasks on at least one previous occasion to ensure good interrater reliability (76).

Validity. *Content validity.* To establish content validity, repeated administrations of the FI from patients with adult PM/DM were analyzed for floor and ceiling effects as well as for internal redundancy and consistency. No tasks were redundant, but grip strength, neck flexion, and trunk flexion (sit-up) showed poor internal consistency with other upper extremity tasks. These results were discussed with a group of health professionals and patients, and hip abduction, transfers, and peak expiratory flow were removed due to ceiling affects and lower relevance. Despite ceiling effects and poor intra- and interrater reliability, the neck flexion was considered relevant and remains in the tool.

Construct validity. The shoulder flexion task correlated moderately with the shoulder flexion isokinetic muscle endurance test (Spearman's r = 0.58) and less with other measures, confirming that the FI-2 assesses muscle endurance in patients with adult PM/DM (76). The knee extension task of the FI-2 (step test) correlated moderately with maximal isokinetic strength of the knee extensors (Spearman's r = 0.42), less with other constructs, and not at all with the isokinetic knee extension endurance test (76). This lack of correlation could be because the step test is performed in a closed-chain movement that also stresses the cardiovascular system, whereas the isokinetic test is open chained.

Criterion validity. There is no gold standard by which to assess criterion validity.

Ability to detect change. Statistically significant improvements were detected in the shoulder flexion task on the right and left sides after a 7-week intensive training program in adult patients with chronic PM/DM (78), with standardized response means between 0.20 and 1.01 for the different components of the FI-2. This study also reported clinically relevant improvements of at least 20% in several of the FI-2 tasks.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The FI-2 assesses muscle endurance, which seems to be an important limitation for patients with PM/DM (81). There is good content validity and reliability and moderate construct validity in patients with adult PM/DM.

Caveats and cautions. The FI-2 takes a rather long time to perform, and further research is needed to establish sensitivity and specificity to change after rehabilitation interventions or medical treatment. Like all other measures of function in myositis, the FI-2 does not discriminate between activity and damage and may diminish in sensitivity and specificity as an activity measure for patients who are farther along in their illness course with accumulated damage and progressive muscle atrophy. Clinical experience indicates that there may be floor effects for several tasks of the FI-2 when used in patients with IBM, although this needs formal evaluation. The FI-2 has been formally tested in adult PM/DM patients, used in IBM patients clinically but not reported on, and has not yet been tested in juvenile myositis.

Clinical usability. While the tool has good psychometric properties in patients with PM/DM, the clinical meaning of scores has not yet been established, making this tool difficult to apply to the care of individual patients. If physical therapists or other personnel are not available to perform the test, the length of time needed to perform the FI-2 is a limiting factor for clinical use. Therefore, a streamlined version of the FI-2 is being developed.

Research usability. The FI-2 has sound content and construct validity and reliability properties in patients with adult PM/DM. The extended numbers of repetitions confirm that the FI-2 assesses muscle endurance, although it was not proven for the knee extension task, which correlated best with isokinetic muscle strength. Additional studies on sensitivity to change and specificity and application to other subgroups of myositis are needed.

MYOSITIS ACTIVITIES PROFILE (MAP)

Description

Purpose. To assess disease-specific limitations of activities of daily living in patients with polymyositis (PM)/dermatomyositis (DM).

Content. The MAP includes 4 subscales (movement activities, activities of moving around, personal care, and domestic activities) and 4 single items (keep in touch with close friends and relatives; avoid overexertion during daily activities; be able to cope with work, studies, and/or housework to a satisfactory degree; and be able to do recreational activities of choice) (24). Subscales and single items are based on the activity domain of the revised International Classification of Impairments, Disability and Handicaps (ICIDH-2) Beta-2 draft (82).

Number of items. The MAP includes 31 items.

Response/option scale. Each item is scored on a 7-point Likert scale from 1–7, where 1 = no trouble to do and 7 = impossible to do.

Recall period for items. During the last week.

Endorsements. None.

Examples of use. The MAP was developed for patients with adult PM/DM and is currently used in clinical practice to evaluate changes after rehabilitation interventions in several rheumatology clinics in Sweden. It is also used in yearly followup visits at the Karolinska University Hospital. The MAP has been used in 1 clinical exercise study.

Practical Application

How to obtain. The MAP can be obtained in English at no cost in the original publication (24) or in Swedish or English by contacting the author (e-mail: helene.alexanderson@karolinska.se) at the Karolinska Institute, Stockholm, Sweden.

Method of administration. The MAP is a self-administered questionnaire.

Scoring. The 4 subscales are scored as the median value of item responses within the subscale. For the subscales movement activities (n = 8 items), moving around (n = 4 items), and domestic activities (n = 6 items), the median value is the lower of the 2 middle values. The subscale personal care (n = 9 items) is scored as the median value. The 4 single items are scored as the actual item response value. In case of missing values that result in an odd number of items in a subscale, the score is the middle value. In case of missing values resulting in an even number of items, the subscale is scored as the lower of the 2 middle values.

Score interpretation. 1 = no difficulty to do and 10 = impossible to do. No cut points have been identified, and normative data are not available.

Respondent burden. The MAP takes 5–10 minutes to complete, with low item difficulty.

Administrative burden. The MAP takes 5 minutes to score by hand.

Translations/adaptations. The MAP has been translated from Swedish into American and British English, and adaptations to the North American and British cultural contexts are ongoing. Only patients with adult PM/DM have been studied to date.

Psychometric Information

Method of development. The items and subscales of the MAP were developed based on the revised ICIDH-2 Beta-2 draft published in 1999 (82). The activity domain of the ICIDH-2 Beta-2 draft included 315 activities classified into the following 8 categories: activities of learning and applying knowledge, communication activities, movement, activities of moving around, self-care activities, domestic activities, interpersonal activities, and performing tasks and major life activities. Eighty-one of these activities from the 6 latter categories were considered by the research group to be relevant for individuals living in Europe. Items were discussed within the research group, and strategically chosen patients with different sexes, diagnoses, disease activity and durations, family situations, and working statuses were invited to rate both the difficulty and importance of items. Ten strategically chosen patients (cohort 1) rated difficulty and importance of the 81 items on a 10-cm visual analog scale (VAS). Questions about sexual activities were rated as limited and very important by cohort 1, but a majority of patients in cohort 2 who filled out the MAP for analysis of internal redundancy and consistency chose not to fill out these questions. Therefore, questions about sexual activities were removed, and the 4 remaining items were listed as single items (24).

Acceptability. Before completing the MAP, patients are asked to decide both how difficult each activity is to perform in daily life and how important it is to be able to perform the activity in daily life. No study to evaluate whether patients can weigh both aspects equally has been carried out. Missing values are rare. No floor or ceiling effects have been detected in the Swedish context (24).

Reliability. *Internal consistency.* Ten strategically chosen patients with adult PM/DM (cohort 1) rated the difficulty and importance of the 81 items on a 10-cm VAS. Spearman's correlation coefficients ranged between 0.61 and 0.91 in testing the internal consistency of subscales (24). There was poor internal consistency between items in the interpersonal activities and performing major life activities subscales.

Test–retest reliability. Weighted κ coefficients for test–retest reliability ranged between 0.56 and 0.76 for subscales and between 0.65 and 0.77 for single items without systematic variations in 17 stable adult PM/DM patients (24).

Validity. *Content validity.* See above in Method of development.

Construct validity. The third version of the MAP correlated highly with the Health Assessment Questionnaire (Spearman's rank correlation = 0.70), but correlated moderately with measures of muscle impairment (Spearman's r = 0.55) and well-being (Spearman's r = 0.43), and poorly with global disease activity (Spearman's r = 0.17) in patients with adult PM/DM (24). Moderate correlations (Spearman's rank correlation = 0.51–0.71) were found between the MAP subscales and single items and the subscales of the Arthritis Impact Measurement Scale (24).

Criterion validity. There is no gold standard by which to establish criterion validity in activity limitation measures.

Ability to detect change. The Swedish MAP has been used as a measure of activity limitation in a 7-week intensive resistance training study that did not reveal statistically significant changes on a group level after short-term exercise therapy in adult PM/DM patients (78). The standardized response mean ranged between 0.15 and 1.32 for the subscales and between 0.20 and 0.41 for the single items.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The MAP is a disease-specific measure of daily life functions, including aspects of both difficulty of the task and importance of the activity. Patients were involved in the development of the tool. There is moderate reliability and moderate construct validity in patients with adult PM/DM.

Caveats and cautions. The MAP needs to be translated to other languages and adapted to other cultural contexts before it is used in clinical practice and research. Information on sensitivity to change and specificity is very limited, and data currently exist for adult PM/DM but not other myositis subgroups. Its applicability to children has also not been examined. Like all other measures of function in myositis, the MAP does not discriminate between activity and damage, and may diminish in sensitivity and specificity as an activity measure for patients who are further along in their illness course with accumulated damage and progressive muscle atrophy. There are no data on the MAP in patients with inclusion body myositis or juvenile idiopathic inflammatory myopathy.

Clinical usability. The low patient and administrative burden and ensured item relevance support its use in clinical practice in Sweden, but the limited language and adaptation availability as well as the lack of cut points and error of measurement are important limitations. The clinical meaning of scores has not yet been established, making this tool difficult to apply to the care of individual patients.

Research usability. The thorough content validity process supports the relevance of items of the MAP, the construct validity analysis shows that the MAP assesses activity limitation, and the acceptable test–retest reliability supports the use of the MAP in research in patients with adult PM/DM. Further research is needed to establish sensitivity to change and specificity and to examine the performance of the MAP in other subgroups of myositis patients.

INCLUSION BODY MYOSITIS FUNCTIONAL RATING SCALE (IBMFRS)

Description

Purpose. The IBMFRS is a 10-point disease-specific functional rating scale that is intended only for patients with inclusion body myositis (IBM) (74).

Content. Includes swallowing, handwriting, cutting food and handling utensils, fine motor tasks, dressing, hygiene, turning in bed and adjusting covers, changing position from sitting to standing, walking, and climbing stairs.

Number of items. 10 items.

Response options/scale. Graded on a Likert scale from 0 (being unable to perform) to 4 (normal).

Recall period for items. Patients are asked to compare how they are at the time the questions are being asked to how they were prior to the start of the disease.

Endorsements. None.

Examples of use. Currently there are 2 clinical trials of interferon (10,60) and an ongoing phase II trial of arimoclomol in IBM that are using the IBMFRS as an outcome measure.

Practical Application

How to obtain. It is available in the original publication and in a review on IBM (74,83).

Method of administration. Interviewer to patient.

Scoring. 10 individual scores are added for a total score.

Score interpretation. Score range is from 0–40, where 40 = normal function and no disability and 0 = severe functional disability. The range of scores corresponding to mild and moderate disability scores has not been determined.

Respondent burden. 15 minutes.

Administrative burden. 15 minutes.

Translations/adaptations. Available in English only. Translations and cross-cultural adaptations are not available. This rating scale has been tested only in patients with IBM, not adult or juvenile polymyositis (PM)/dermatomyositis (DM).

Psychometric Information

Method of development. It was modified from the Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS), which was developed to allow patients to rate their muscle function, self-care, and pulmonary function (84).

Acceptability. Missing data are not common because it is a set of questions that is asked. If a question is missed, the scores available would be added. There are no floor or ceiling effects.

Reliability. A reliability study for IBM is in progress.

Validity. *Content validity.* The instrument was developed by neurologists, clinical evaluators, and the research coordinators in the Muscle Study Group. The ALSFRS was used as the template, and several items were altered to address motor problems specific to IBM patients.

Construct validity. The IBMFRS showed significant moderate to good correlations (Pearson's correlation coefficients 0.55–0.86) with maximal voluntary isometric contraction, Manual Muscle Testing, handgrip dynamometry, and the ALSFRS in IBM patients (74).

Criterion validity. There are no criterion validity results available for the IBMFRS.

Ability to detect change. This instrument was shown to be able to detect change in a 24-week trial of interferon-β for IBM, with an effect size of −2.9 (74).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument does measure important elements of functional disability for patients with IBM. It is quick, inexpensive, and easy to administer. It does not require any special equipment or training.

Caveats and cautions. The clinician asks the patient to compare how they are today with how they were before the start of the disease. Some IBM patients have had the disease for decades. For them it might be harder to remember their state before disease onset. Also, as people get older, they tend to lose function in the hand or get arthritis (harder to use keys, pick up objects). It can be difficult to separate normal aging processes from IBM-related processes. Like all other measures of function in myositis, the IBMFRS does not discriminate between activity and damage, and may diminish in sensitivity and specificity as an activity measure for patients who are farther along in their illness course with accumulated damage and progressive muscle atrophy. Further validation of the IBMFRS is needed, particularly for patients with IBM. The IBMFRS has not been developed for or tested in patients with adult or juvenile PM/DM.

Clinical usability. The IBMFRS should be a valuable clinical tool, since it is quick and easy to administer.

Research usability. It is easily incorporated into IBM research protocols. It is the only IBM-specific outcome measure based on subject responses.

CUTANEOUS DERMATOMYOSITIS DISEASE AREA AND SEVERITY INDEX (CDASI)

Description

Purpose. The CDASI is a clinician- or clinician–investigator-scored instrument that separately measures activity and damage in the skin of dermatomyositis (DM) patients (7,85). Because it is a 1-page instrument with common and responsive elements, it is feasible to use in daily clinical practice for monitoring DM skin disease. There is a modified CDASI (version 2), which is the one in current use

(85). This modified version further simplifies the original CDSAI by combining ulceration and erosion into 1 category, simplifies descriptors for Gottron's damage and nailfold changes, and eliminates excoriation as a subscale (85).

Content. The modified CDASI has 3 activity measures (erythema, scale, and erosion/ulceration) and 2 damage measures (poikiloderma and calcinosis). In addition, Gottron's papules on the hands are evaluated in terms of activity (erythema, ulceration) and damage (dyspigmentation or scarring). Lastly, activity in terms of periungual changes and alopecia is measured.

Number of items and subscales. Each of the 3 activity scales (erythema, scale, and erosion/ulceration) and 2 damage measures (poikiloderma and calcinosis) is assessed over 15 body areas; the worst level of activity is scored, whereas the damage measures are scored for their presence or absence. In additon, Gottron's papules are evaluated in terms of activity (erythema or ulceration) and damage (dyspigmentation or scarring). Lastly, activity in terms of periungual changes and alopecia is measured.

Response options/scale. Disease activity is assessed by the worst degree of erythema (1 = pink, 2 = red, 3 = dark red), scale (1 = scale, 2 = crust, lichenification), and the presence of erosions or ulceration (scored as present or absent) in 15 different anatomic locations. Periungual changes are scored from 0–2, where 0 = no periungual changes, 1 = periungual erythema, and 2 = visible telangiectasias. Alopecia is scored present or absent, where 0 = no alopecia and 1 = presence of alopecia in the past 30 days. Gottron's sign on the knuckles is assessed similarly to the erythema scale used in other anatomic locations. When Gottron's papules are present, the erythema score obtained on the knuckles is doubled. Disease damage is assessed by the presence or absence of poikiloderma or calcinosis in the 15 different anatomic locations. In additon, damage in areas of Gottron's sign on the hands is assessed (1 = dyspigmentation, 2 = scarring).

Recall period for items. Current examination, except for alopecia that may be present over the past 30 days.

Endorsements. None.

Examples of use. The modified CDASI has been used in several prospective databases of adult DM patients and in 2 completed therapeutic trials.

Practical Application

How to obtain. The CDASI is copyrighted and can only be reprinted with permission from the authors. The CDASI may be used for routine clinical use by clinicians in order to assist the clinical consultation, evaluation, and clinical decision-making process. There is no need to seek specific permission for this and there is no charge for the use of the CDASI in this context. However, it is a requirement that every copy of the CDASI should always reprint the copyright statement: "© University of Pennsylvania 2009." There is a requirement to seek permission when the CDASI is used for research purposes. Purely academic research projects are granted use of the CDASI without charge. Please contact Dr. Victoria Werth (e-mail: werth@mail.med.upenn.edu) for permission to use.

Method of administration. The CDASI is administered by a trained clinician while examining the patient.

Scoring. Each item of the CDASI version 2 is scored according to the most severe lesions in a body area and on the various characteristics outlined above. The CDASI has a total score ranging from 0–132, which is divided into activity and damage subscores, which range from 0–100 and 0–32, respectively. Scoring of disease activity, as indicated on the CDASI instrument, involves adding the scores on the left half of the CDASI, i.e., erythema, scale, erosion/ulcerations, Gottron's sign, periungual change, and alopecia. Scoring of disease damage requires addition of scores on the right half of the CDASI, i.e., poikiloderma, calcinosis, and Gottron's dyspigmentation or scarring. Missing values are counted as 0.

Score interpretation. Scores range from 0–100 for activity and from 0–32 for damage. Among the activity items, the potential range for erythema for all 15 areas is 0–45, for scale is 0–30, and for erosion/ulcerations is 0–15. The range for Gottron's erythema is 0–6, for Gottron's ulcerations is 0–6, for periungual change is 0–2, and for alopecia is 0–1. For damage items, the range for poikiloderma is 0–15, for calcinosis is 0–15, and for Gottron's damage is 0–2. Higher scores indicate greater disease activity or greater disease damage.

The level of disease activity can be interpreted as low, moderate, or high. The mean ± SD CDASI activity for mild disease was 11.4 ± 7.0, moderate was 25.6 ± 8.9, and severe was >39.4 (86). Ongoing studies are refining mild, moderate, and severe disease categories and examining the minimal clinically significant change. Scores in other populations are not available but presumably would be 0 for a healthy individual.

Respondent burden. Not applicable.

Administrative burden. The CDASI takes a mean of 4.8 minutes for dermatologists experienced in the assessment of dermatomyositis to complete (7); presumably less experienced physicians may take longer. Training is necessary for reliable assessment of activity and damage. A training tool is available from Dr. Werth. Scoring takes <1 minute and can be done by hand.

Translations/adaptations. The CDASI is available in English. It has been studied and used in patients with adult classic DM, as well as with hypomyopathic and amyopathic DM, but not in other myositis subgroups.

Psychometric Information

Method of development. Development of the CDASI has been an iterative process involving experts in rheumatologic dermatology. The CDASI was designed to capture the most important signs of activity and damage that are predominant in patients with DM and signs that would be amenable to change over time (7,85). Dermatologists experienced in the assessment of DM thought that the CDASI was complete, and they expressed satisfaction with the measure during multi-investigator meetings and studies. Items were generated by discussion of important aspects of the disease with patients and by discussion of specific items with expert dermatologists during group meetings. Subscales were generated based on items chosen by the group during consensus meetings as important measures of cutaneous DM activity and damage, with elements of activity selected as responsive to change. The tool was modified due to the group's desire to simplify the CDASI and to better describe some of the elements of the subscales.

Acceptability. The instrument is 1 page and easily readable. Missing data are not common, and any missing items are scored as 0. Data analyzed from a prospective database of 182 dermatomyositis assessments have not shown floor or ceiling effects.

Reliability. *Internal consistency.* Evidence for internal consistency is not currently available.

Test–retest reliability. The CDASI had an intraclass correlation coefficient (ICC) for the CDASI activity subscore of 0.84 (95% confidence interval [95% CI] 0.70–0.98) (7). The CDASI had an ICC for the CDASI damage subscore of 0.86 (95% CI 0.75–0.98) (7). Intrarater reliability ICC for the modified CDASI activity subscore was 0.87 (95% CI 0.70–0.95), and for the modified CDASI damage subscore was 0.80 (95% CI 0.56–0.92) (85).

Interrater reliability. The CDASI had an ICC for the CDASI activity subscore of 0.84 (95% CI 0.70–0.98) (7). The CDASI had an ICC for the CDASI damage subscore of 0.53 (95% CI 0.32–0.73) (7). Interrater reliability ICC for the modified CDASI activity subscore was 0.75 (95% CI 0.55–0.90), and for the modified CDASI damage subscore was 0.56 (95% CI 0.36–0.79) (85).

Validity. *Content validity.* Evaluation of content was considered adequate by all 10 dermatologists participating in a validation study with the modified CDASI. Content validation is described further above in Method of development.

Construct validity. The physician global activity (Spearman's r = 0.75) and damage visual analog scales (Spearman's r = 0.90) correlate highly with the activity and damage subscores of the modified CDASI (85), and a global itch score correlates moderately (Spearman's r = 0.63) with the CDASI activity score from a study of adult DM (85). CDASI activity scores correlated moderately (Pearson's r = 0.46 for emotion, 0.44 for function, and 0.33 for symptoms) with the Skindex-29 subscores and correlated mildly but significantly with the Dermatology Life Quality Index (r = 0.29) in patients with adult DM, suggesting that increased cutaneous activity, as measured by the CDASI, correlates with a poorer quality of life.

Criterion validity. The CDASI was found to be a significant predictor of the Likert scales for physician global activity and damage scores, which were the compared gold standards. All CDASI mean scores (total, activity, and damage) expressed statistically significant distinct values when grouped by Likert scores (mild, moderate, severe activity or damage, all P values ≤ 0.001) (7). The CDASI expressed a significant, near-perfect fit for linearity for activity ($P < 0.001$) and damage ($P < 0.005$), with r^2 values ≥ 0.95 (87).

Responsiveness to change. CDASI scores were assessed, as well as a physician global score and an overall evaluation from the physician, as to whether the patient had improved, worsened, or not changed from their previous research visit. The standardized response mean (SRM) for the largest clinical change per patient, defined as the largest difference in the physician global activity score between 2 consecutive visits, was 1.25 for the CDASI, which corresponded to an SRM of 1.03 for physician global activity (87).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CDASI is a partially validated 1-page instrument that captures key findings regarding skin activity and damage in DM patients. It allows capture of the worst attributes of 15 body areas but does not involve measurement of body surface area (BSA). BSA is notoriously difficult to capture, particularly for a condition that may involve only small amounts of skin. The tool attempts to assess improvement within an area by providing several levels of activity for erythema, scale, and Gottron's lesions. A small modification to simplify the CDASI was shown to have equally good validity and reliability in comparison with the original CDASI. Currently the CDASI shows good reliability, good but limited construct validity, and excellent responsiveness in patients with adult DM.

Caveats and cautions. Appropriate training on use of the CDASI is suggested to reduce variability in assessments. Definition and measurement of poikiloderma often involve a component of erythema and dyspigmentation, both of which are captured. Further studies of the CDASI are needed to determine cut points for mild, moderate, and severe skin disease activity and damage, as well as the minimal clinically significant change needed to demonstrate improvement. The instrument was designed to measure important responsive elements but was not designed to capture every element of DM skin disease. The CDASI has been used and partially validated in adult DM patients, but not in other subgroups of myositis.

Clinical usability. Based on available psychometric data, the CDASI should be a useful measure in the clinical context. Calculation is simple, with separate determination of a total activity and a total damage score for an overall score by simply adding them. This separation of activity and damage scores prevents the potential for paradoxical stability of scores as disease activity decreases, but damage simultaneously worsens.

Research usability. The CDASI has been useful in research assessments. The CDASI has been used in several multicenter studies to evaluate response in the skin of DM patients. Studies looking at response to therapy will likely focus on the CDASI activity assessment, which has been shown to be responsive to change.

CUTANEOUS ASSESSMENT TOOL (CAT)

Description

Purpose. The CAT was developed to comprehensively assess a wide range of cutaneous manifestations of idiopathic inflammatory myopathy (IIM) in children and adults (88). It was first published in 2007. An abbreviated version of the CAT (aCAT) was published in 2008 and is currently the preferred format (89).

Content. Items of the CAT were chosen by expert opinion to reflect the range of both activity and damage in cutaneous lesions observed in juvenile and adult IIM.

Number of items. The CAT consists of a skin disease activity score and a skin disease damage score. There are a total of 21 items, including 10 activity lesions, 4 damage lesions, and 7 lesions that are common to both the activity and damage scores.

Response options/scale. In the original CAT, each lesion is scored depending on various characteristics (e.g., erythema, scaling). For the aCAT, each item is either present or absent.

Recall period for items. Scoring of the CAT requires that the lesion be observed at the time the CAT is administered (i.e., no recall period).

Endorsements. None.

Examples of use. The CAT has been used to date in studies that have examined its psychometric properties (7,87–90).

Practical Application

How to obtain. The CAT is available from the *Rheumatology* web site (posted as supplementary material) (88) and on the International Myositis Assessment and Clinical Studies Group web site (http://www.niehs.nih.gov/research/resources/collab/imacs/othertools.cfm).

Method of administration. The CAT is administered by a trained clinician while examining the patient.

Scoring. Each item of the CAT is scored depending on the presence of the lesion and on various characteristics (e.g., erythema, presence of scaling, crusting, or erosions, and presence of ulcerations or necrosis). Scores for each item range from 0–2 to 0–7. For the aCAT, items are scored as 1 if present and 0 if absent.

Score interpretation. For the original CAT, the total skin disease activity score ranges from 0–96, and the total skin disease damage score ranges from 0–20. For the aCAT, the total skin disease activity score ranges from 0–17, and the total skin disease damage score ranges from 0–11. A score of 0 reflects the absence of cutaneous manifestations. When compared to a 5-point ordinal scale for disease activity and damage, median (25th to 75th percentiles) CAT activity scores, corresponding to "no evidence of skin disease activity," "mild," "moderate," "severe," and "very severe skin disease activity," were 1 (0–3), 7 (4–9), 13 (10–20), 18 (12–33), and 31 (27–39), respectively. The median (25th to 75th percentiles) CAT damage scores, corresponding to "no evidence of skin disease damage," "mild," "moderate," and "severe or very severe skin disease damage," were 0 (0–1), 1 (0–2), 2 (1–4), and 5 (3–6), respectively (90).

Respondent burden. Depending on the complexity of skin disease of a patient, the CAT takes up to 15 minutes to complete, although 1 study using dermatologists experienced in the assessment of dermatomyositis skin disease reported a mean of 5 minutes (7). The aCAT takes less time due to the removal of detailed scoring.

Administrative burden. The time it takes to administer the CAT may be a limitation in a busy clinic. Scoring takes <1 minute and can be done by hand. Training in the administration of the CAT is preferred.

Translations/adaptations. None available at present. The CAT has been studied and partially validated in juvenile polymyositis/dermatomyositis (DM) patients and adult DM patients.

Psychometric Information

Values of psychometric evaluations for the aCAT were nearly identical to those for the CAT (89).

Method of development. The development of the CAT was undertaken by a group of adult and pediatric rheumatologists and a pediatric dermatologist (88). Items were chosen based on expert opinion regarding the important cutaneous lesions of IIM. Twenty-eight lesions were considered candidates, including 16 activity lesions, 5 damage lesions, and 7 lesions that represented a combination of activity and damage. This list was reviewed by a larger group of rheumatologists and dermatologists, resulting in the deletion of 5 lesions (purpura, Raynaud's phenomenon, urticaria, mucinous papules, and acanthosis nigricans) and the combination of 4 other lesions into 2 lesions (Gottron's papules with Gottron's sign, malar erythema with facial erythema). Scoring was determined by the investigators based on consensus expert opinion (88).

Acceptability. Given that the tool is administered by the clinician, missing data are not common. Missing data are scored as 0 or absent. The length of the tool has been criticized (hence development of the aCAT).

Reliability. *Internal consistency.* When juvenile IIM patients were assessed by pediatric rheumatologists, the standardized Cronbach's α for the CAT activity score was 0.79. Individual standardized Cronbach's α scores ranged from 0.77–0.81 when each item was removed from the activity score. The standardized Cronbach's α for the CAT damage score was 0.74. Individual standardized Cronbach's α scores ranged from 0.67–0.76 when each item was removed from the damage score (90). Item-total correlations for the CAT ranged from 0.02–0.67 for the activity items and from 0.001–0.29 for the damage items. The items with low correlations were generally those present in few patients, and they improved to a minimum of 0.27 ($P \leq 0.05$) for lesions with >10% endorsement. Item to domain correlations for the activity items ranged from 0.25–0.99 and increased to a minimum of 0.42 ($P \leq 0.05$) for lesions with ≥10% endorsement (90). Internal consistency of the aCAT was comparable to the full CAT, with Cronbach's α of 0.76 for the aCAT activity score and 0.70 for the aCAT damage score (89).

Test–retest reliability. In adult patients with IIM assessed by dermatologists, the CAT activity score had an intraclass correlation coefficient (ICC) of 0.74 (95% confidence interval [95% CI] 0.50–0.95), and the CAT damage score had an ICC of 0.58 (95% CI 0.27–0.89) (7).

Interrater reliability. This was assessed by having assessors review images of typical IIM lesions. ICCs for each lesion ranged from 0.33–0.90 (90). In juvenile IIM patients seen by 2 assessors, ICCs for the total activity and total damage scores were 0.71 and 0.81, respectively. ICCs for the individual items ranged from 0.11–1.0 (88). ICCs for the aCAT were comparable (0.60 for the total aCAT activity and 0.65 for total aCAT damage) (89). In a study of adults with DM assessed by dermatologists experienced in DM, the CAT activity score had an ICC of 0.60 (95% CI 0.40–0.79), and the CAT damage score had an ICC of 0.43 (95% CI 0.22–0.64) (7). The ICC for the aCAT was 0.55 (87).

Validity. *Content validity.* This has not been formally reassessed in IIM since the original development of this tool.

Construct validity. For children with juvenile IIM assessed by pediatric rheumatologists, the CAT activity score correlated highly with the 10-cm visual analog scale (VAS) for physician skin disease activity (Spearman's r = 0.83, $P < 0.0001$) and physician global disease activity (Spearman's r = 0.77, $P < 0.0001$), and moderately with measures of muscle strength and function (correlation with Childhood Myositis Assessment Scale = −0.48, with Childhood Health Assessment Questionnaire = 0.40, and with total Manual Muscle Testing = −0.36) (90). As expected, the CAT activity and damage scores correlated poorly with serum levels of muscle enzymes (Spearman's r = 0.03–0.13), but the CAT activity score correlated mildly but significantly with lactate dehydrogenase (0.37) (90). The CAT damage score correlated moderately with the 10-cm VAS for physician skin disease damage (Spearman's r = 0.53, $P < 0.0001$) and for physician global disease damage (Spearman's r = 0.52, $P <$ 0.0001) (90).

In adult patients with DM assessed by dermatologists, the CAT activity score had a Spearman's r of −0.69 with the physician global disease activity and a Spearman's r of −0.53 with 10-cm VAS for patient global disease activity. Correlation with the global itch score was moderate (Spearman's r = 0.59). The CAT damage score had a Spearman's r of −0.47 with 10-cm VAS for physician disease damage and a Spearman's r of −0.13 for 10-cm VAS for patient disease damage (7). The aCAT was also found to correlate significantly with the physician global activity VAS in a study of adult DM patients (87).

When the scores were evaluated in relation to levels of physician global activity in adult DM patients, the patients with mild global disease activity had mean ± SD CAT scores of 8.3 ± 5.1, patients with moderate global activity had mean ± SD CAT scores of 15.2 ± 6.9, and patients with severe disease activity had mean ± SD CAT scores of 22.5 ± 7.4 (7).

Criterion validity. There is no gold standard by which to establish criterion validity.

Ability to detect change. In children with juvenile IIM assessed 7–9 months apart, the standardized response mean (SRM) of the CAT activity score was 0.52 (95% CI 0.32–0.72). In children with a >0.8-cm improvement in physician skin disease activity, the SRM was 0.67 (95% CI 0.42–0.92) (90). SRM values for the CAT damage score were not relevant over the duration of this study. In adult DM patients, the SRM was 0.93 in a group of patients who had exhibited change based on a physician's rating (87).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The CAT and aCAT are comprehensive measures that assess the full range of cutaneous lesions in IIM. The requirement that the patient being assessed is observed reduces the likelihood of biases in reporting. The CAT and aCAT have good reliability, construct validity, and responsiveness in patients with juvenile (ages 2–18 years) and adult DM.

Caveats and cautions. Appropriate training is preferred to reduce variability in assessments. There are some concerns about the reliability of some items. The tool has been partially validated in juvenile IIM and adult DM, but not examined in other myositis subgroups.

Clinical usability. Based on available psychometric data, the CAT and aCAT should be useful measures in the clinical context. The time needed to administer the full CAT may be a limitation for clinicians.

Research usability. The CAT and aCAT should be useful in research assessments. The lack of information concerning change over time in the CAT damage score should lead to some caution if used for this purpose.

DERMATOMYOSITIS SKIN SEVERITY INDEX (DSSI)

Description

Purpose. The DSSI assesses disease activity in skin of dermatomyositis (DM) patients. The tool is patterned after the Psoriasis Area and Severity Index (PASI) (91).

Content. The DSSI assesses disease activity based on involved body surface area (BSA) and severity. Body area is divided into 4 parts (head, trunk, upper extremity, and lower extremity) and scored by percentage of involvement. Severity of involvement is scored for the 4 anatomic locations with 3 symptom scores (redness, induration, and scaliness). The DSSI is calculated based on the percentage BSA involved (92).

Number of items and subscales. Each of these 4 body areas is assessed by visual inspection for redness, induration, and scaliness.

Response options/scale. The areas involved in each of the 4 main body areas are measured on the following 0–6-point scale: 0 = no involvement, 1 = <10% involvement, 2 = 10–30% involvement, 3 = 31–50% involvement, 4 = 51–70% involvement, 5 = 71–90% involvement, and 6 = 91–100% involvement. The average redness, induration, and scaliness of the lesions in each of the body areas are scored on a 0–4-point scale (91).

Recall period for items. Current examination. There is no recall period.

Endorsements. None.

Practical Application

How to obtain. This tool is available at no cost and is published (91). E-mail Dr. Joseph Jorizzo (jjorizzo@wfubmc.edu) for permission to use.

Method of administration. The DSSI is administered by a trained clinician while examining the patient.

Scoring. The sum of the redness, induration, and scaliness scores (maximum of 12) is multiplied by the area score for each body area (maximum of 6). These totals are normalized (10%, 20%, 30%, and 40% for the head, upper extremities, trunk, and lower extremities, respectively) and summed. The total DSSI score can range from 0–72, with higher scores representing more severe disease activity (92). There are no instructions for missing values, but these are presumably scored as 0.

Score interpretation. When compared to the global physician activity score, the mean ± SD DSSI scores were 1.3 ± 1.5 for mild global activity, 5.4 ± 4.0 for moderate global activity, and 14.9 ± 14.1 for severe global disease activity (7).

Respondent burden. Not applicable**.**

Administrative burden. Completion takes ~2–3 minutes for experienced dermatologists who are familiar with the tool. Training is needed, as done for the PASI in psoriasis, and can be accessed on the following web site: http://www.pasitraining.com/index.html.

Translations/adaptations. The DSSI is available in English. It has been validated and studied in patients with adult DM, but not other myositis subgroups.

Psychometric Information

Method of development. Initial content of the scale was validated for content by a panel of experts that included board-certified dermatologists and rheumatologists. This score is mirrored after the PASI.

Acceptability. Given that the tool is administered by a clinician, missing data are not common. The tool is rapid to use.

Reliability. *Internal consistency.* Internal consistency has not been statistically evaluated.

Stability. Test–retest stability has been evaluated, with intraclass correlation coefficients (ICCs) between examinations by the same observers ranging from 0.79 (95% confidence interval [95% CI] 0.34–0.95) to 0.93 (95% CI 0.87–0.99) (91,92).

Intrarater reliability. Intrarater reliability has been completed in adult DM or amyopathic DM, ranging from 0.79 (95% CI 0.34–0.95) to 0.89 (95% CI 0.76–0.95) (7,92).

Interrater reliability. The DSSI has been tested at 3 institutions, with ICCs ranging from 0.44 (95% CI 0.23–0.65) to 0.94 (95% CI 0.84–0.97) in patients with adult DM or amyopathic DM (7,92).

Validity. *Content validity.* Content validity was evaluated by a panel of expert dermatologists and rheumatologists and found to be adequate (91).

Construct validity. The DSSI correlates moderately with physician global disease activity (Spearman's r = 0.51–0.83) and also with pruritis (Spearman's r = 0.41–0.61) in adult DM patients (7,92). The DSSI was also found to correlate moderately with the presence of poikiloderma (Spearman's r = 0.61–0.70) (91), although the DSSI is supposed to measure activity, and poikiloderma is typically associated with damage. In evaluation of quality of life relative to the DSSI, the Spearman's correlations were also moderate (Spearman's r = 0.41 with the Skindex-16 and 0.38 with the Dermatology Life Quality Index) in adult DM patients (92). There was no significant correlation between the DSSI and periungual capillary nailfold changes, cutaneous ulceration, calcinosis, muscle enzyme levels, or muscle strength (92).

Criterion validity. There is no gold standard upon which to assess criterion validity.

Responsiveness to change. In 1 study of adult DM patients who received a variety of treatments, the DSSI showed a mean change of 3.9 units after treatment (95% CI 1.0–6.9). The Spearman's correlation coefficient between the change in DSSI scores and the change in physician global activity was 0.28 (92). Additional evaluation of the responsiveness of the DSSI is not available.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. This tool is a measure of skin disease activity in DM and is based on another scale, the PASI, which has been used widely in psoriasis therapeutic trials. The measure is quick to use by experienced dermatologists. The measure has acceptable reliability and limited but moderate construct validity in patients with adult DM and amyopathic DM.

Caveats and cautions. The DSSI is a disease activity measure that depends on assessment of BSA based on the rule of 9s. BSA can be difficult to assess reliably, particularly when only small areas are involved, as can occur in DM (93). Responsiveness to change when small areas of skin are involved will likely be difficult using a measure that depends on BSA. The DSSI does not include an assessment of damage. The tool has been used in patients with adult amyopathic and classic DM, but not in other subgroups of myositis.

Clinical usability. The DSSI is easy to use, but psychometric properties suggest that it might be difficult to use accurately. There are no measurements of damage.

Research usability. The usability for research depends on how extensive the disease process is. It may be difficult to demonstrate change in patients with limited BSA involvement. There is no measurement of damage.

SKINDEX

Description

Purpose. To measure quality of life (QOL) in different populations and detect changes over time. This is a clinically responsive measure for the effect of skin disease on patients' QOL (94–96). It has been used in acne, psoriasis, atopic dermatitis, seborrheic dermatitis, alopecia areata, vitiligo, nevi, skin cancer, cutaneous lupus, and dermatomyositis (DM), among other skin conditions (97). There are several versions, with the Skindex-29 the most utilized and validated. Initially, the Skindex was a 61-item self-administered survey that measured cognitive effects, social effects, depression, fear, embarrassment, anger, physical discomfort, and physical limitations (94). It has been modified and refined several times. The tool was shortened to 29 items, with the same reliability and validity, but with more discriminative and evaluative features (95). In 2001, the Skindex-16 was published (98). It is a sensitive, accurate, single-page survey and has 2 additional advantages compared with the Skindex-29 (98). It evaluates the most bothersome rather than the most frequent symptoms, and it has fewer items, due to less duplication of questions where most patients choose the same response. A Skindex-17 is also available, developed using Rasch analysis (99). There is experience with the Skindex-16 and Skindex-29 in patients with DM.

Content. For the Skindex-29, each item is scored on a 5-point Likert scale: 0 = never, 1 = rarely, 2 = sometimes, 3 = often, and 4 = all the time. For the Skindex-16, each item is scored on a scale of 1 (never bothered) to 7 (always bothered). Both tools have 3 subscales (emotion, symptoms, and functioning).

Number of items and subscales. The Skindex-29 has 30 items, 29 of which are used for scoring. Three questions were added to represent DM-specific effects, specifically 2 questions for photosensitivity and 1 question for alopecia. All responses are transformed to a linear scale of 100, varying from 0 (no effect) to 100 (effect experienced all the time). Each question and subscale ranges from 0–100 points, with higher scores indicating worse QOL.

Recall period for items. 4 weeks.

Endorsements. None.

Practical Application

How to obtain. There is no cost; e-mail Dr. Mary-Margaret Chren (mchren@orca.ucsf.edu) for permission to use and guidance about scoring (95).

Method of administration. Self-administered questionnaire.

Scoring. Individual items (1–5) are added to yield a total score for each subscale; higher scores indicate worse QOL. A composite score has not been formally studied, has no face validity, and did not fit the Rasch model (99).

Score interpretation. Norms, as well as correlation with QOL burden in a number of different skin diseases, are available (100). For the Skindex-29, an additional study evaluated patients with a mix of diseases, with >60% of the patients having an inflammatory skin disease such as acne, psoriasis, or seborrheic dermatitis, and almost one-half of the patients graded as having at least moderate disease severity, to determine the clinical meaning of scores according to symptom severity for each of the subscales (100). This study demonstrated that the emotions subscale had a mean of 3.2 (cut point of <5) for very little disease, a mean of 16 (cut points of 6–24) for mild impact on emotions, a mean of 36.6 (cut points of 25–49) for moderate, and a mean of 62.6 (cut point of >50) for severe emotional impact. The symptoms subscale had a mean of 0.0 (cut points of 0–3) for very little symptoms, a mean of 6.6 (cut points of 4–10) for mild, a mean of 17.6 (cut points of 11–25) for moderate, a mean of 37.3 (cut points of 26–49) for severe, and a mean of 62.2 (cut point of >50) for extremely severe symptoms. The function subscale showed a mean of 0.0 (cut point of <3) for very little functional impairment, a mean of 5.3 (cut points of 4–10) for mild, a mean of 20.6 (cut points of 11–32) for moderate, and a mean of 48.6 (cut point of >33) for severe functional impairment (100).

Respondent burden. It takes ~5 minutes for patients to complete the questionnaire.

Administrative burden. Time for scoring is <1 minute.

Translations/adaptations. The Skindex is available in English, Spanish, Dutch, German, French, Italian, Arabic, and Turkish. To date, it has been studied in patients with many different skin diseases, including adult DM and amyopathic DM, as well as inflammatory, autoimmune, and other skin conditions (97,101).

Psychometric Information

Method of scoring. All responses are transformed to a linear scale of 100, varying from 0 (no effect) to 100 (effect experienced all the time). Therefore, each item can have a minimum score of 0 and a maximum score of 100. A scale score is the mean of a patient's responses to the items in a given scale. If responses to >25% of items are missing, the questionnaire is eliminated in research settings. If any scale has >25% of the responses missing, the scale is eliminated. Scale scores are the average of items in a given scale (no imputation). A composite score is defined as the average of all items in the instrument. Any patient for whom all 3 scales are missing should be eliminated from the analytic data set.

Acceptability. This is easy to read, missing data are not common, and there have not been floor or ceiling effects with the diseases studied.

Reliability. *Internal consistency.* For the Skindex-29, the Cronbach's α was 0.87–0.96 for dermatology patients with a mix of inflammatory skin diseases (52%), skin cancers, and benign lesions (95). The Skindex-16 exhibited good internal consistency for each of the scales (Cronbach's α = 0.86, 0.93, and 0.92 for the symptoms, emotions, and functioning scales, respectively) in patients with adult DM (101).

Test–retest reliability. Skindex scale scores were reproducible after 72 hours (r = 0.88–0.92) when tested in a subset of dermatology outpatients (95). The Skindex-16 shows similar reliability in patients with DM and amyopathic DM (101).

Interrater reliability. This has not been evaluated for patients with DM.

Validity. *Content validity.* The initial Skindex-61 items and scales were generated from literature review and focus sessions with dermatology patients, physicians, and nurses. The Skindex-29 items and scales were derived from the Skindex-61 by means of psychometric analysis (95). Three additional items related to photosensitivity and alopecia were added to the Skindex-29. Content validity has not been formally assessed for DM.

Criterion validity. In a study of a variety of dermatology patients, this scale differentiated between skin diseases presumed to have high impact and skin diseases presumed to have a low impact (95). When the Skindex-29 subscores were used to compare adult DM with other dermatologic diseases, DM had among the highest mean subscores, with the emotional subscore being among the most severely affected in patients with DM. DM also showed a higher mean symptom subscore than most compared groups and had a significantly higher score compared to patients with other inflammatory skin conditions, as well as those with normal skin.

Concurrent validity. Evidence of convergent validity is provided by the pattern of correlation between the Skindex and Short Form 36 (SF-36) comparative scales. For each comparative scale, patients in tertiles classified by low, medium, or high responses to Skindex differed according to scores in the corresponding SF-36 comparative scales (96). In adult DM, the emotional subscale of the Skindex correlated moderately well with 3 emotional subscales of the SF-36.

Construct validity. Skindex scores correlated more highly than SF-36 scores with patients' self-reports of the condition of their skin and their perceived disfigurement from their skin disease (96). Each of the Skindex-29 subscores significantly correlated with the Dermatology Life Quality Index scores (Skindex-29 symptom r = 0.63–0.86) (101). Skindex subscores correlated mildly to moderately with Cutaneous Dermatomyositis Disease Area and Severity Index scores (r = 0.32–0.46) in adult DM and amyopathic DM patients. A global pruritus visual analog scale (VAS) correlated moderately with Skindex symptoms and function (Spearman's r = 0.46–0.60) and poorly with Skindex emotion (Spearman's r = 0.19). In evaluation of QOL relative to the Dermatomyositis Skin Severity Index, the correlation was moderate (Spearman's r = 0.41) in adult DM and amyopathic DM patients (92). Pruritus VAS correlated moderately (Spearman's r = 0.60) in patients with adult DM (101). Three emotional subscores of the SF-36 moderately correlated with the emotional subscore of the Skindex. As expected, the Health Assessment Questionnaire, a measure of general physical disability, does not correlate well with the emotional scale of the Skindex.

Responsiveness to change. Mean scale scores remained stable or changed appropriately in patients with a variety of dermatologic conditions over a 3-month period (97). Responsiveness is not available for patients with DM.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The Skindex-29 captures skin-specific QOL issues and corresponds to the severity of skin disease in DM, as well as other skin diseases. A general QOL measure like the SF-36 correlates more highly with increasing degrees of comorbidity and worse self-reported health status, but focuses on functional limitations and emotional state regardless of cause. The Skindex correlates more highly than the SF-36 scores with patients' reports of the condition of their skin. The Skindex is particularly good for evaluating the emotional component of QOL relative to some other measures available. The Skindex (−16 and −29) has internal consistency, test–retest reliability data, and moderate construct validity in patients with adult DM and amyopathic DM.

Caveats and cautions. This questionnaire is longer than some other skin-specific QOL measures. The meaning of the composite score is less clear than the subscale scores, and scores for subscales are used most frequently. Several items show item bias across sex, age, disease severity, and diagnosis (99). The tool to date has no data on responsiveness, and has not been studied in other myositis subgroups.

Clinical usability. Based on available psychometric data, the Skindex should be a useful measure in the clinical context. It has been used in many different skin diseases and has been carefully validated, but validity in myositis is limited.

Research usability. The Skindex has been useful in research assessments of skin diseases, including in 1 study of patients with DM. Further studies of the validity in patients with myositis are needed.

DERMATOLOGY LIFE QUALITY INDEX (DLQI)

Description

Purpose. The DLQI, developed in 1994, was the first dermatology-specific quality of life (QOL) instrument (102). It is a simple, compact, and practical questionnaire for use in dermatology clinical settings to assess QOL in skin disease. Although the DLQI covers a wide range of life impairments, it is not a multiple-scale questionnaire; its scoring system is restricted to an overall score. There are 2 versions of the DLQI for adults and 2 versions for children: a text-only version and an illustrated version. The illustrated version of the DLQI has been shown to correlate with the text-only version (103). The text version has been used in numerous studies, including the assessment of cutaneous disease as part of other autoimmune diseases, as well as in the evaluation of inflammatory and noninflammatory skin conditions (102,104–106) There is a children's version of the DLQI, the Children's DLQI, with a text and cartoon version, the latter of which is preferred by children (107,108).

Content. The measure consists of 10 questions encompassing skin symptoms and feelings, daily activities, leisure, work or school, personal relationships, and the side effects of treatment.

Number of items and subscales. 10 items, no subscales.

Response options/scale. Each item is scored on a Likert scale, where 0 = not at all/not relevant, 1 = a little, 2 = a lot, and 3 = very much.

Recall period for items. 1 week.

Endorsements. None.

Practical Application

How to obtain. The DLQI has been published (102,103); the developer was Dr. Andrew Y. Finlay, Department of Dermatology, Cardiff University School of Medicine, Wales, UK (e-mail: FinlayAY@cf.ac.uk). The DLQI may be used by any clinician worldwide for routine clinical practice without seeking permission and without charge. For details of other uses of the DLQI, including research studies, see http://www.dermatology.org.uk/quality/quality-life.html.

Method of administration. Self-administered questionnaire. The cartoon version has been used for children as young as age 4 years. Parents may complete a parent version of the questionnaire.

Scoring. Scores of individual items (0–3) are added to yield a total score (0–30); higher scores mean greater impairment of patients' QOL as impacted by their skin disease.

Score interpretation. Cut points have been determined for scores, corresponding to 0 (score of 0–1, no effect), 1 (score of 2–5, small effect), 2 (score of 6–10, moderate effect), 3 (score of 11–20, very large effect), and 4 (score of 21–30, extremely large effect) in a questionnaire study involving a number of different inflammatory, malignant, and other skin conditions (107).

Respondent burden. Time for answering the questionnaire is an average of ~2 minutes.

Administrative burden. Scoring takes <1 minute. No training is needed for scoring.

Translations/adaptations. The DLQI is available in 55 languages (104). The DLQI has been studied in the diseases mentioned in the descriptive section above. To date, it has undergone limited study in adults with amyopathic and classic dermatomyositis (DM), but not in other myositis subgroups.

Psychometric Information

Method of development. Initially, 120 patients generated a list of the ways in which their lives were affected by their skin diseases. This led to identification of 49 aspects of QOL impairments, generating a 10-item questionnaire that was subsequently modified slightly, followed by pilot testing in additional patients (102,104). This instrument was developed in the UK with patients visiting a university clinic; it focused on patients' ability to function in their daily activities and does not fully capture emotions and mental health (97).

Acceptability. The DLQI is very readable and easy to complete. Missing data are uncommon. Floor effects have been seen with certain items related to everyday activities and the work/study dimension (109). There are also substantial ceiling effects, with 2 items contributing to most of the variability of the DLQI (109–111).

Reliability. *Internal consistency.* Cronbach's α for the DLQI was assessed in patients with a variety of skin conditions, and ranged from 0.75–0.92 (104). This has not been assessed in DM.

Test–retest reliability. Test–retest reliability of the DLQI has ranged from 0.56–0.99 in patients with a variety of skin conditions. Most studies showed values >0.90 (104). This has not been assessed in DM.

Validity. *Content validity.* Content validity was established by examining the ability of the instrument to discriminate between patients with skin disease and normal healthy subjects ($P < 0.001$) (109). There is a question of content related to emotion in adult DM, where the emotional component of QOL is extremely important. Specifically, the correlation between DLQI and Skindex-29 function scores were significantly higher than the correlation between DLQI and Skindex-29 emotion scores in adult DM patients ($P = 0.004$).

Construct validity. The DLQI has been used in many studies that have shown significant correlation between the DLQI and generic, dermatology-specific, and disease-specific measures (104). There is low to moderate correlation (Spearman's r = 0.36–0.38) of the DLQI with the Dermatomyositis Skin Severity Index in DM and amyopathic DM patients (92). There is moderate to excellent correlation of the DLQI with Skindex-29 subscores (Pearson's r = 0.63–0.86) in DM and amyopathic DM patients (111). The DLQI exhibited significant but poor correlation with the Cutaneous Dermatomyositis Disease Area and Severity Index (Pearson's r = 0.35) and with a global pruritus visual analog scale (VAS; Pearson's r = 0.27). However, in a second study of adult DM patients, there was moderate correlation of the DLQI with a global pruritus VAS (Spearman's r = 0.58) (101).

Correlations between the DLQI and other dermatology-specific health-related QOL measures were high (r = 0.6–0.86), were moderate for general health-related QOL measures (r = 0.3–0.62), and were in the expected directions except that the DLQI correlates less with mental and emotional aspects (97,105). Concurrent correlation with the Short Form 36 was demonstrated in an acne study (r = −0.44 to −0.33) (104). In adult DM patients, the DLQI correlated better with the Skindex function subscale (r = 0.86) relative to the Skindex symptoms subscale (r = 0.63) or emotion subscale (r = 0.67).

Criterion validity. The cut points of the DLQI using global questions show a κ of 0.489 (112). This has not been assessed in DM.

Responsiveness to change. The ability to detect small impairments may be difficult because of substantial ceiling effects (109–111). However, many studies have demonstrated responsiveness to change (104). The minimum clinically important difference of the DLQI in specific skin diseases has been estimated to range from 2.2–6.9, based on data from 5 studies in other skin diseases (104). Information on the responsiveness and minimum clinically important difference does not exist for DM.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The DLQI focuses on the impact of skin diseas on patients' ability to function in their daily activities and might not fully capture emotions and mental health (106,111). The strength of the DLQI is its simplicity and broad use for clinical investigation in dermatology, with application to a number of skin conditions (104,105). The DLQI has limited moderate construct validity in adult DM and amyopathic DM.

Caveats and cautions. There has been concern that emotions and mental health can be very important in inflammatory skin diseases such as DM. One study found that in DM the correlations between DLQI and Skindex-29 function scores were significantly higher than the correlation between DLQI and Skindex-29 emotion scores ($P =$

0.004), suggesting that the DLQI might not capture the full range of emotional QOL. There are several limitations related to the focus on disability, response distribution, and dimensionality and item bias. To date, there are no studies of its reliability or responsiveness in adult DM, and no studies in other subgroups of myositis.

Clinical usability. The DLQI has been used in numerous studies and trials of a number of skin conditions, although it is limited in its study in adult DM patients. It is clinically easy to use.

Research usability. The DLQI has been well evaluated for a variety of skin diseases and works well for research, with the caveat that the emotional aspect of QOL may be captured better with other instruments. It is thought to be unidimensional, with scoring restricted to an overall score. Validation data in adult DM are limited.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Miller FW, Rider LG, Chung YL, Cooper R, Danko K, Farewell V, et al. Proposed preliminary core set measures for disease outcome assessment in adult and juvenile idiopathic inflammatory myopathies. Rheumatology (Oxford) 2001;40:1262–73.
2. Ruperto N, Ravelli A, Pistorio A, Ferriani V, Calvo I, Ganser G, et al, for the Paediatric Rheumatology International Trials Organisation (PRINTO) and the Pediatric Rheumatology Collaborative Study Group (PRCSG). The provisional Paediatric Rheumatology International Trials Organisation/American College of Rheumatology/European League Against Rheumatism disease activity core set for the evaluation of response to therapy in juvenile dermatomyositis: a prospective validation study. Arthritis Rheum 2008;59:4–13.
3. Rider LG, Giannini EH, Brunner HI, Ruperto N, James-Newton L, Reed AM, et al, for the International Myositis Assessment and Clinical Studies Group. International consensus on preliminary definitions of improvement in adult and juvenile myositis. Arthritis Rheum 2004; 50:2281–90.
4. Ruperto N, Pistorio A, Ravelli A, Rider LG, Pilkington C, Oliveira S, et al, for the Paediatric Rheumatology International Trials Organisation (PRINTO) and the Pediatric Rheumatology Collaborative Study Group (PRCSG). The Paediatric Rheumatology International Trials Organisation provisional criteria for the evaluation of response to therapy in juvenile dermatomyositis. Arthritis Care Res (Hoboken) 2010;62: 1533–41.
5. Rider LG, Feldman BM, Perez MD, Rennebohm RM, Lindsley CB, Zemel LS, et al, in cooperation with the Juvenile Dermatomyositis Disease Activity Collaborative Study Group. Development of validated disease activity and damage indices for the juvenile idiopathic inflammatory myopathies. I. Physician, parent, and patient global assessments. Arthritis Rheum 1997;40:1976–83.
6. Rider LG, Giannini EH, Harris-Love M, Joe G, Isenberg D, Pilkington C, et al. Defining clinical improvement in adult and juvenile myositis. J Rheumatol 2003;30:603–17.
7. Klein RQ, Bangert CA, Costner M, Connolly MK, Tanikawa A, Okawa J, et al. Comparison of the reliability and validity of outcome instruments for cutaneous dermatomyositis. Br J Dermatol 2008;159:887–94.
8. Ruperto N, Ravelli A, Murray KJ, Lovell DJ, Andersson-Gare B, Feldman BM, et al. Preliminary core sets of measures for disease activity and damage assessment in juvenile systemic lupus erythematosus and juvenile dermatomyositis. Rheumatology (Oxford) 2003;42:1452–9.
9. Harris-Love MO, Shrader JA, Koziol D, Pahlajani N, Jain M, Smith M, et al. Distribution and severity of weakness among patients with polymyositis, dermatomyositis and juvenile dermatomyositis. Rheumatology (Oxford) 2009;48:134–9.
10. Muscle Study Group. Randomized pilot trial of BINF1a (Avonex) in patients with inclusion body myositis. Neurology 2001;57:1566–70.
11. Rider LG, Koziol D, Giannini EH, Jain MS, Smith MR, Whitney-Mahoney K, et al. Validation of manual muscle testing and a subset of eight muscles for adult and juvenile idiopathic inflammatory myopathies. Arthritis Care Res (Hoboken) 2010;62:465–72.
12. Sanner H, Kirkhus E, Merckoll E, Tollisen A, Roisland M, Lie BA, et al. Long-term muscular outcome and predisposing and prognostic factors in juvenile dermatomyositis: a case–control study. Arthritis Care Res (Hoboken) 2010;62:1103–11.
13. Jain M, Smith M, Cintas H, Koziol D, Wesley R, Harris-Love M, et al. Intra-rater and inter-rater reliability of the 10-point Manual Muscle Test (MMT) of strength in children with juvenile idiopathic inflammatory myopathies (JIIM). Phys Occup Ther Pediatr 2006;26:5–17.
14. Isenberg DA, Allen E, Farewell V, Ehrenstein MR, Hanna MG, Lundberg IE, et al. International consensus outcome measures for patients with idiopathic inflammatory myopathies: development and initial validation of myositis activity and damage indices in patients with adult onset disease. Rheumatology (Oxford) 2004;43:49–54.
15. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
16. Clarke AE, Bloch DA, Medsger TA Jr, Oddis CV. A longitudinal study of functional disability in a national cohort of patients with polymyositis/dermatomyositis. Arthritis Rheum 1995;38:1218–24.
17. Mercer LK, Moore TL, Chinoy H, Murray AK, Vail A, Cooper RG, et al. Quantitative nailfold video capillaroscopy in patients with idiopathic inflammatory myopathy. Rheumatology (Oxford) 2010;49:1699–705.
18. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
19. Neri R, Mosca M, Stampacchia G, Vesprini E, Tavoni A, d'Ascanio A, et al. Functional and isokinetic assessment of muscle strength in patients with idiopathic inflammatory myopathies. Autoimmunity 2006;39:255–9.
20. Singh G, Athreya BH, Fries JF, Goldsmith DP. Measurement of health status in children with juvenile rheumatoid arthritis. Arthritis Rheum 1994;37:1761–9.
21. Huber AM, Hicks JE, Lachenbruch PA, Perez MD, Zemel LS, Rennebohm RM, et al, for the Juvenile Dermatomyositis Disease Activity Collaborative Study Group. Validation of the Childhood Health Assessment Questionnaire in the juvenile idiopathic myopathies. J Rheumatol 2001;28:1106–11.
22. Feldman BM, Ayling-Campos A, Luy L, Stevens D, Silverman ED, Laxer RM. Measuring disability in juvenile dermatomyositis: validity of the Childhood Health Assessment Questionnaire. J Rheumatol 1995;22:326–31.
23. Ponyi A, Borgulya G, Constantin T, Vancsa A, Gergely L, Danko K. Functional outcome and quality of life in adult patients with idiopathic inflammatory myositis. Rheumatology (Oxford) 2005;44:83–8.
24. Alexanderson H, Lundberg IE, Stenstrom CH. Development of the Myositis Activities Profile: validity and reliability of a self-administered questionnaire to assess activity limitations in patients with polymyositis/dermatomyositis. J Rheumatol 2002;29:2386–92.
25. Takken T, Elst E, Spermon N, Helders PJ, Prakken AB, van der Net J. The physiological and physical determinants of functional ability measures in children with juvenile dermatomyositis. Rheumatology (Oxford) 2003;42:591–5.
26. Maillard SM, Jones R, Owens C, Pilkington C, Woo P, Wedderburn LR, et al. Quantitative assessment of MRI T2 relaxation time of thigh muscles in juvenile dermatomyositis. Rheumatology (Oxford) 2004; 43:603–8.
27. Rider LG, Lachenbruch PA, Monroe JB, Ravelli A, Cabalar I, Feldman BM, et al. Damage extent and predictors in adult and juvenile dermatomyositis and polymyositis as determined with the Myositis Damage Index. Arthritis Rheum 2009;60:3425–35.
28. Apaz MT, Saad-Magalhaes C, Pistorio A, Ravelli A, de Oliveira Sato J, Marcantoni MB, et al, for the Paediatric Rheumatology International Trials Organisation. Health-related quality of life of patients with juvenile dermatomyositis: results from the Paediatric Rheumatology International Trials Organisation multinational quality of life cohort study. Arthritis Rheum 2009;61:509–17.
29. Ruperto N, Ravelli A, Pistorio A, Malattia C, Cavuto S, Gado-West L, et al. Cross-cultural adaptation and psychometric evaluation of the Childhood Health Assessment Questionnaire (CHAQ) and the Child Health Questionnaire (CHQ) in 32 countries: review of the general methodology. Clin Exp Rheumatol 2001;19 Suppl:S1–9.
30. Huber AM, Feldman BM, Rennebohm RM, Hicks JE, Lindsley CB, Perez MD, et al, for the Juvenile Dermatomyositis Disease Activity Collaborative Study Group. Validation and clinical significance of the Childhood Myositis Assessment Scale for assessment of muscle function in the juvenile idiopathic inflammatory myopathies. Arthritis Rheum 2004;50:1595–603.
31. Lovell DJ, Lindsley CB, Rennebohm RM, Ballinger SH, Bowyer SL, Giannini EH, et al, in cooperation with the Juvenile Dermatomyositis Disease Activity Collaborative Study Group. Development of validated disease activity and damage indices for the juvenile idiopathic inflammatory myopathies. II. The Childhood Myositis Assessment

Scale (CMAS): a quantitative tool for the evaluation of muscle function. Arthritis Rheum 1999;42:2213–9.

32. Rennebohm RM, Jones K, Huber AM, Ballinger SH, Bowyer SL, Feldman BM, et al, for the Juvenile Dermatomyositis Disease Activity Collaborative Study Group. Normal scores for nine maneuvers of the Childhood Myositis Assessment Scale. Arthritis Rheum 2004;51:365–70.
33. Huber AM, Giannini EH, Bowyer SL, Kim S, Lang B, Lindsley CB, et al. Protocols for the initial treatment of moderately severe juvenile dermatomyositis: results of a Children's Arthritis and Rheumatology Research Alliance Consensus Conference. Arthritis Care Res (Hoboken) 2010;62:219–25.
34. Ravelli A, Trail L, Ferrari C, Ruperto N, Pistorio A, Pilkington C, et al. Long-term outcome and prognostic factors of juvenile dermatomyositis: a multinational, multicenter study of 490 patients. Arthritis Care Res (Hoboken) 2010;62:63–72.
35. Whiting-O'Keefe QE, Stone JH, Hellmann DB. Validity of a vasculitis activity index for systemic necrotizing vasculitis. Arthritis Rheum 1999;42:2365–71.
36. Hay EM, Bacon PA, Gordon C, Isenberg DA, Maddison P, Snaith ML, et al. The BILAG index: a reliable and valid instrument for measuring clinical disease activity in systemic lupus erythematosus. Q J Med 1993;86:447–58.
37. Sultan SM, Allen E, Oddis CV, Kiely P, Cooper RG, Lundberg IE, et al. Reliability and validity of the Myositis Disease Activity Assessment Tool. Arthritis Rheum 2008;58:3593–9.
38. Stone KB, Oddis CV, Fertig N, Katsumata Y, Lucas M, Vogt M, et al. Anti–Jo-1 antibody levels correlate with disease activity in idiopathic inflammatory myopathy. Arthritis Rheum 2007;56:3125–31.
39. Symmons DP, Coppock JS, Bacon PA, Bresnihan B, Isenberg DA, Maddison P, et al. Development and assessment of a computerized index of clinical disease activity in systemic lupus erythematosus. Q J Med 1988;69:927–37.
40. Cresswell L, Yee CS, Farewell V, Rahman A, Teh LS, Griffiths B, et al. Numerical scoring for the classic BILAG index. Rheumatology (Oxford) 2009;48:1548–52.
41. Bode RK, Klein-Gitelman MS, Miller ML, Lechman TS, Pachman LM. Disease Activity Score for children with juvenile dermatomyositis: reliability and validity evidence. Arthritis Rheum 2003;49:7–15.
42. Smith RL, Sundberg J, Shamiyah E, Dyer A, Pachman LM. Skin involvement in juvenile dermatomyositis is associated with loss of end row nailfold capillary loops. J Rheumatol 2004;31:1644–9.
43. Rouster-Stevens KA, Morgan GA, Wang D, Pachman LM. Mycophenolate mofetil, a possible therapeutic agent for children with juvenile dermatomyositis. Arthritis Care Res (Hoboken) 2010;62:1446–51.
44. Schmeling H, Stephens S, Goia C, Manlhiot C, Schneider R, Luthra S, et al. Nailfold capillary density is importantly associated over time with muscle and skin disease activity in juvenile dermatomyositis. Rheumatology (Oxford) 2011;50:885–93.
45. Sultan SM, Ioannou Y, Moss K, Isenberg DA. Outcome in patients with idiopathic inflammatory myositis: morbidity and mortality. Rheumatology (Oxford) 2002;41:22–6.
46. Sadjadi R, Rose MR. What determines quality of life in inclusion body myositis? J Neurol Neurosurg Psychiatry 2010;81:1164–6.
47. Regardt M, Welin HE, Alexanderson H, Lundberg IE. Patients with polymyositis or dermatomyositis have reduced grip force and health-related quality of life in comparison with reference values: an observational study. Rheumatology (Oxford) 2011;50:578–85.
48. Alexanderson H, Stenstrom CH, Lundberg I. Safety of a home exercise programme in patients with polymyositis and dermatomyositis: a pilot study. Rheumatology (Oxford) 1999;38:608–11.
49. Alexanderson H, Stenstrom CH, Jenner G, Lundberg I. The safety of a resistive home exercise program in patients with recent onset active polymyositis or dermatomyositis. Scand J Rheumatol 2000;29:295–301.
50. Landgraf JM, Abetz L, Ware JE. The CHQ user's manual. 1st ed. Boston: The Health Institute, New England Medical Center; 1996.
51. Ruperto N, Martini A. International research networks in pediatric rheumatology: the PRINTO perspective. Curr Opin Rheumatol 2004;16:566–70.
52. Martini A, Ruperto N, for the Paediatric Rheumatology International Trials Organisation (PRINTO). Quality of life in juvenile idiopathic arthritis patients compared to healthy children. Clin Exp Rheumatol 2001;19 Suppl:S1–172.
53. Duffy CM, Arsenault L, Duffy KN, Paquin JD, Strawczynski H. The Juvenile Arthritis Quality of Life Questionnaire: development of a new responsive index for juvenile rheumatoid arthritis and juvenile spondyloarthritides. J Rheumatol 1997;24:738–46.
54. Varni JW, Seid M, Knight TS, Burwinkle T, Brown J, Szer IS. The PedsQL in pediatric rheumatology: reliability, validity, and responsiveness of the Pediatric Quality of Life Inventory Generic Core Scales and Rheumatology Module. Arthritis Rheum 2002;46:714–25.
55. Stoll T, Seifert B, Isenberg DA. SLICC/ACR Damage Index is valid, and renal and pulmonary organ scores are predictors of severe outcome in patients with systemic lupus erythematosus. Br J Rheumatol 1996;35:248–54.
56. Rahman P, Gladman DD, Urowitz MB, Hallett D, Tam LS. Early damage as measured by the SLICC/ACR Damage Index is a predictor of mortality in systemic lupus erythematosus. Lupus 2001;10:93–6.
57. Sultan SM, Allen E, Cooper RG, Agarwal S, Kiely P, Oddis CV, et al. Interrater reliability and aspects of validity of the Myositis Damage Index. Ann Rheum Dis 2011;70:1272–6.
58. Sanner H, Gran JT, Sjaastad I, Flato B. Cumulative organ damage and prognostic factors in juvenile dermatomyositis: a cross-sectional study median 16.8 years after symptom onset. Rheumatology (Oxford) 2009;48:1541–7.
59. Mathiesen PR, Zak M, Herlin T, Nielsen SM. Clinical features and outcome in a Danish cohort of juvenile dermatomyositis patients. Clin Exp Rheumatol 2010;28:782–9.
60. Muscle Study Group. Randomized pilot trial of high-dose β INF1a in patients with inclusion body myositis. Neurology 2004;63:718–20.
61. Barohn RJ, Herbelin L, Kissel JT, King W, McVey AL, Saperstein DS, et al. Pilot trial of etanercept in the treatment of inclusion-body myositis. Neurology 2006;66 Suppl:S123–4.
62. Rutkove SB, Parker RA, Nardin RA, Connolly CE, Felice KJ, Raynor EM. A pilot randomized trial of oxandrolone in inclusion body myositis. Neurology 2002;58:1081–7.
63. Dalakas MC, Sonies B, Dambrosia J, Sekul E, Cupler E, Sivakumar K. Treatment of inclusion-body myositis with IVIg: a double-blind, placebo-controlled study. Neurology 1997;48:712–6.
64. Dalakas MC, Koffman B, Fujii M, Spector S, Sivakumar K, Cupler E. A controlled study of intravenous immunoglobulin combined with prednisone in the treatment of IBM. Neurology 2001;56:323–7.
65. Brussock CM, Haley SM, Munsat TL, Bernhardt DB. Measurement of isometric force in children with and without Duchenne's muscular dystrophy. Phys Ther 1992;72:105–14.
66. Rose MR, McDermott MP, Thornton CA, Palenski C, Martens WB, Griggs RC. A prospective natural history study of inclusion body myositis: implications for clinical trials. Neurology 2001;57:548–50.
67. Stoll T, Bruhlmann P, Stucki G, Seifert B, Michel BA. Muscle strength assessment in polymyositis and dermatomyositis evaluation of the reliability and clinical use of a new, quantitative, easily applicable method. J Rheumatol 1995;22:473–7.
68. Tawil R, McDermott MP, Mendell JR, Kissel J, Griggs RC. Facioscapulohumeral muscular dystrophy (FSHD): design of natural history study and results of baseline testing. Neurology 1994;44:442–6.
69. The FSH-DY Group. A prospective, quantitative study of the natural history of facioscapulohumeral muscular dystrophy (FSHD): implications for therapeutic trials. Neurology 1997;48:38–46.
70. Meldrum D, Cahalane E, Conroy R, Fitzgerald D, Hardiman O. Maximum voluntary isometric contraction: reference values and clinical application. Amyotroph Lateral Scler 2007;8:47–55.
71. Stoll T, Huber E, Seifert B, Michel BA, Stucki G. Maximal isometric muscle strength: normative values and gender-specific relation to age. Clin Rheumatol 2000;19:105–13.
72. Escolar DM, Henricson EK, Mayhew J, Florence J, Leshner R, Patel KM, et al. Clinical evaluator reliability for quantitative and manual muscle testing measures of strength in children. Muscle Nerve 2001;24:787–93.
73. Personius KE, Pandya S, King WM, Tawil R, McDermott MP. Facioscapulohumeral dystrophy natural history study: standardization of testing procedures and reliability of measurements. Phys Ther 1994;74:253–63.
74. Jackson CE, Barohn RJ, Gronseth G, Pandya S, Herbelin L. Inclusion body myositis functional rating scale: a reliable and valid measure of disease severity. Muscle Nerve 2008;37:473–6.
75. Dalakas MC, Rakocevic G, Schmidt J, Salajegheh M, McElroy B, Harris-Love MO, et al. Effect of alemtuzumab (CAMPATH 1-H) in patients with inclusion-body myositis. Brain 2009;132:1536–44.
76. Alexanderson H, Broman L, Tollback A, Josefson A, Lundberg IE, Stenstrom CH. Functional Index-2: validity and reliability of a disease-specific measure of impairment in patients with polymyositis and dermatomyositis. Arthritis Rheum 2006;55:114–22.
77. Josefson A, Romanus E, Carlsson J. A functional index in myositis. J Rheumatol 1996;23:1380–4.
78. Alexanderson H, Dastmalchi M, Esbjornsson-Liljedahl M, Opava CH, Lundberg IE. Benefits of intensive resistance training in patients with chronic polymyositis or dermatomyositis. Arthritis Rheum 2007;57:768–77.
79. Borg GA. Psychophysical bases of perceived exertion. Med Sci Sports Exerc 1982;14:377–81.
80. Nordenskiold UM, Grimby G. Grip force in patients with rheumatoid arthritis and fibromyalgia and in healthy subjects: a study with the Grippit instrument. Scand J Rheumatol 1993;22:14–9.
81. Alexanderson H, Bergegard J, Alemo-Munters L, Dastmalchi M, Lundberg IE. Patients with idiopathic inflammatory myopathies have low

muscle endurance rather than low muscle strength [abstract]. Arthritis Rheum 2009;60 Suppl:S307.

82. World Health Organization. International classification of impairment, disability and handicap: β-2 draft, full version. Geneva: World Health Organization; 1999.
83. Amato AA, Barohn RJ. Inclusion body myositis: old and new concepts. J Neurol Neurosurg Psychiatry 2009;80:1186–93.
84. The ALS CNFT Treatment Study (ACTS) Phase I-II Study Group. The Amyotrophic Lateral Sclerosis Functional Rating Scale: assessment of activities of daily living in patients with amyotrophic lateral sclerosis. Arch Neurol 1996;53:141–7.
85. Yassaee M, Fiorentino D, Okawa J, Taylor L, Coley C, Troxel AB, et al. Modification of the Cutaneous Dermatomyositis Disease Area and Severity Index, an outcome instrument. Br J Dermatol 2010;162:669–73.
86. Chock M, Goreshi R, Werth VP, Fiorentino D. Quantitative assessment of disease severity in dermatomyositis [abstract]. J Invest Dermatol 2010;130 Suppl:S50.
87. Goreshi R, Okawa J, Rose M, Feng R, Lee LA, Hansen C. Evaluation of reliability, validity, and responsiveness of the CDASI and CAT-BM [abstract]. Arthritis Rheum 2010;62 Suppl:S381.
88. Huber AM, Dugan EM, Lachenbruch PA, Feldman BM, Perez MD, Zemel LS, et al. The Cutaneous Assessment Tool: development and reliability in juvenile idiopathic inflammatory myopathy. Rheumatology (Oxford) 2007;46:1606–11.
89. Huber AM, Lachenbruch PA, Dugan EM, Miller FW, Rider LG, for the Juvenile Dermatomyositis Disease Activity Collaborative Study Group. Alternative scoring of the Cutaneous Assessment Tool in juvenile dermatomyositis: results using abbreviated formats. Arthritis Rheum 2008;59:352–6.
90. Huber AM, Dugan EM, Lachenbruch PA, Feldman BM, Perez MD, Zemel LS, et al, in cooperation with the Juvenile Dermatomyositis Disease Activity Collaborative Study Group. Preliminary validation and clinical meaning of the Cutaneous Assessment Tool in juvenile dermatomyositis. Arthritis Rheum 2008;59:214–21.
91. Fredriksson T, Pettersson U. Severe psoriasis: oral therapy with a new retinoid. Dermatologica 1978;157:238–44.
92. Carroll CL, Lang W, Snively B, Feldman SR, Callen J, Jorizzo JL. Development and validation of the Dermatomyositis Skin Severity Index. Br J Dermatol 2008;158:345–50.
93. Charman CR, Venn AJ, Williams HC. Measurement of body surface area involvement in atopic eczema: an impossible task? Br J Dermatol 1999;140:109–11.
94. Chren MM, Lasek RJ, Quinn LM, Mostow EN, Zyzanski SJ. Skindex, a quality-of-life measure for patients with skin disease: reliability, validity, and responsiveness. J Invest Dermatol 1996;107:707–13.
95. Chren MM, Lasek RJ, Flocke SA, Zyzanski SJ. Improved discriminative and evaluative capability of a refined version of Skindex, a quality-of-life instrument for patients with skin diseases. Arch Dermatol 1997;133:1433–40.
96. Chren MM, Lasek RJ, Quinn LM, Covinsky KE. Convergent and discriminant validity of a generic and a disease-specific instrument to measure quality of life in patients with skin disease. J Invest Dermatol 1997;108:103–7.
97. Both H, Essink-Bot ML, Busschbach J, Nijsten T. Critical review of generic and dermatology-specific health-related quality of life instruments. J Invest Dermatol 2007;127:2726–39.
98. Chren MM, Lasek RJ, Sahay AP, Sands LP. Measurement properties of Skindex-16: a brief quality-of-life measure for patients with skin diseases. J Cutan Med Surg 2001;5:105–10.
99. Nijsten TE, Sampogna F, Chren MM, Abeni DD. Testing and reducing Skindex-29 using Rasch analysis: Skindex-17. J Invest Dermatol 2006;126:1244–50.
100. Nijsten T, Sampogna F, Abeni D. Categorization of Skindex-29 scores using mixture analysis. Dermatology 2009;218:151–4.
101. Hundley JL, Carroll CL, Lang W, Snively B, Yosipovitch G, Feldman SR, et al. Cutaneous symptoms of dermatomyositis significantly impact patients' quality of life. J Am Acad Dermatol 2006;54:217–20.
102. Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI): a simple practical measure for routine clinical use. Clin Exp Dermatol 1994;19:210–6.
103. Loo WJ, Diba V, Chawla M, Finlay AY. Dermatology Life Quality Index: influence of an illustrated version. Br J Dermatol 2003;148:279–84.
104. Basra MK, Fenech R, Gatt RM, Salek MS, Finlay AY. The Dermatology Life Quality Index 1994-2007: a comprehensive review of validation data and clinical results. Br J Dermatol 2008;159:997–1035.
105. Lewis V, Finlay AY. 10 years experience of the Dermatology Life Quality Index (DLQI). J Investig Dermatol Symp Proc 2004;9:169–80.
106. De Korte J, Mombers FM, Sprangers MA, Bos JD. The suitability of quality-of-life questionnaires for psoriasis research: a systematic literature review. Arch Dermatol 2002;138:1221–7.
107. Lewis-Jones MS, Finlay AY. The Children's Dermatology Life Quality Index (CDLQI): initial validation and practical use. Br J Dermatol 1995;132:942–9.
108. Holme SA, Man I, Sharpe JL, Dykes PJ, Lewis-Jones MS, Finlay AY. The Children's Dermatology Life Quality Index: validation of the cartoon version. Br J Dermatol 2003;148:285–90.
109. Badia X, Mascaro JM, Lozano R. Measuring health-related quality of life in patients with mild to moderate eczema and psoriasis: clinical validity, reliability and sensitivity to change of the DLQI. Br J Dermatol 1999;141:698–702.
110. Morgan M, McCreedy R, Simpson J, Hay RJ. Dermatology quality of life scales: a measure of the impact of skin diseases. Br J Dermatol 1997;136:202–6.
111. Shikiar R, Willian MK, Okun MM, Thompson CS, Revicki DA. The validity and responsiveness of three quality of life measures in the assessment of psoriasis patients: results of a phase II study. Health Qual Life Outcomes 2006;4:71.
112. Hongbo Y, Thomas CL, Harrison MA, Salek MS, Finlay AY. Translating the science of quality of life into practice: what do dermatology life quality index scores mean? J Invest Dermatol 2005;125:659–64.

Summary Table for Measures of Disease Activity in Myositis*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
Physician global activity	Overall rating of myositis disease activity based on all clinical and laboratory measures available at the time of assessment	Clinician completed	N/A	<1 min, but time to assess the patient Hand scored	On 10-cm VAS, 0 = inactive disease, 10 = extremely severe disease activity On a Likert scale, 0 = inactive, 1 = mild activity, 2 = moderate activity, 3 = severe activity, 4 = extremely severe activity	Excellent internal consistency and interrater reliability	Excellent content and construct validity	Excellent responsiveness ≥20% improvement is consensus of clinically meaningful change	Measures important concept Good psychometric properties Appropriate for clinical and research use Most validated in juvenile PM/DM with some validity in adult PM/DM	Somewhat subjective and based on the experience of the rater Reliability of serial ratings dependent on examining previous scores Not validated for IBM
Patient/parent global activity	Overall rating of myositis disease activity	Patient or parent self-report	<1 min	<1 min Hand scored	On 10-cm VAS, 0 = inactive disease, 10 = extremely severe disease activity On a Likert scale, 0 = inactive, 1 = mild activity, 2 = moderate activity, 3 = severe activity, 4 = extremely severe activity	Ratings distinct from physician ratings. Reliability not available for patient/parent global activity	Good construct validity	Excellent responsiveness ≥20% improvement is consensus of clinically meaningful change	Measures important concept Good psychometric properties Appropriate for clinical and research use Most validated in juvenile PM/DM with some validity in adult PM/DM	Somewhat subjective and based on the experience of the rater Reliability of serial ratings dependent on examining previous scores Not validated for IBM
MMT	Measures muscle strength by application of pressure to muscle groups tested against gravity or through a range of motion for muscle groups with less than antigravity strength	Administered by a trained clinician/ physical therapist	Takes 30–60 min to assess 24–26 muscle groups Takes <5 min to assess 8 key muscle groups May be demanding for weak child or younger children with limited ability to cooperate	Takes 30–60 min to assess 24–26 muscle groups Takes <5 min to assess 8 key muscle groups Hand scoring <1 min	Modified MRC or Kendall's 0–10 scales used. Scores may be 0–260 for a total score of 12 proximal and distal muscle groups tested bilaterally + 2 axial muscle groups, 0–80 for MMT8	Excellent internal consistency, test–retest reliability, very good intrarater reliability. Reliability of scores much better than of individual muscle groups	Good content validity for MMT8 Good construct validity	Excellent responsiveness ≥15% improvement in MMT score in adult PM/DM and ≥18% improvement for juvenile IIM is consensus for clinically important improvement	Measures concept central to the assessment of myositis patients Sound psychometric properties Appropriate for clinical and research use Validated in adult and juvenile PM/DM. Does not require special equipment	Requires training in administration of the test Widely used but not validated in IBM Total MMT is lengthy for clinical setting Does not distinguish activity from damage Patients with muscle atrophy may not be sensitive to change
HAQ/C-HAQ	Assesses physical function in 9 (HAQ) or 8 (C-HAQ) domains of daily activities	Self- or proxy administered	Minimal	Minimal	Range 0–3: 0 = no or mild physical dysfunction, <0.125–0.25 = mild physical dysfunction, >1.0 = moderate to severe disability	Test–retest reliability excellent in children Internal consistency acceptable for juvenile myositis Intrarater reliability not available in myositis	Construct validity excellent in children. Some evidence supportive in adult PM/DM No assessment of content validity in myositis Limited criterion validity in adult PM/DM	Responsiveness good to excellent in children with recognized change Data not available for adults	Brief and easy to use Takes little time Good psychometrics in children with myositis	Significant floor effect Limited validity in adult PM/DM No validity in IBM

(continued)

Summary Table *(Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
CMAS	Assesses muscle strength, physical function, and endurance	Observational, performance based, administered by clinician or therapist	15–20 min May be demanding for weak child or younger children with limited ability to cooperate	15–20 min to administer, <1 min to score	Range 0–52 Higher scores indicate greater strength or physical function: <15 = severe weakness (consensus), >48 = normal, >45 = mild impairment, >39 = mild to moderate, >30 = moderate impairment (based on comparison with C-HAQ)	Test–retest and interrater reliability very good to excellent Internal consistency not available	Strong evidence for construct validity Content validity not assessed	Responsiveness strong in children with recognized change	Comprehensive assessment that specifically addresses endurance Reduction in bias and noncompletion due to observational nature Good psychometric properties in juvenile IIM	Requires training to administer Time needed to administer may limit usefulness clinically Significant ceiling effect Currently validated only for juvenile IIM and not studied yet in adult myositis subgroups.
MDAAT	Assesses 6 extramuscular organs to produce a global extramuscular score, and the muscle score, which gives a total disease activity index score	Clinician completed	N/A	Time to complete a history and physical examination (likely 15–30 min) Hand or computer scored	For MYOACT organ system score scored on 10-cm VAS, 0 = inactive disease, 10 = extremely severe disease activity. For MYOACT each item is answered 0 = not present, 1= improving, 2 = same, 3 = worse, 4 = new, and converted to organ system scores of A–E based on the intent-to-treat. Scores range from 0–60 for the extramuscular MYOACT score and 0–70 for the total MYOACT score, and they range from 0–54 for the extramuscular MITAX score and 0–63 for the total MITAX score	Excellent internal consistency, good interrater reliability	Excellent content validity Good construct validity	Excellent responsiveness ≥20% improvement in the extramuscular score is consensus of clinically meaningful change	Measures important concept Good psychometric properties Appropriate for clinical and research use Most validated in adult and juvenile DM	Somewhat cumbersome to use/score (needs training) MYOACT scores are somewhat subjective and based on the experience of the rater For MYOACT scores, reliability of serial ratings dependent on examining previous scores Not validated for IBM
DAS	Evaluates muscle and skin involvement	Clinician completed	5 min	Hand calculated	Total score: 0–20, with higher scores meaning higher disease activity Skin subscale 0–9 Weakness subscale 0–11	Good internal consistency, moderate to poor interrater reliability	Good construct validity	Excellent responsiveness	Simplicity and good psychometric properties Validation studies in juvenile DM	Performance in clinical trials still to be evaluated Not evaluated in other myositis subgroups
SF-36	Assessment of the global HRQOL, functional health, and well-being	Self-administered	Minimal	Minimal Scoring is by computer	Consists of 36 items answered by marking from 2–6 options Scoring ranges from 0–100, with 0 = maximum disability	Test–retest reliability and intrarater reliability are not available in myositis	Good construct and criterion validity in DM and PM, with limited data in IBM Content validity is unavailable in myositis	Statistics on responsiveness not available in myositis	Widely used in other diseases Easily administered Available in multiple languages, with extensive normative data	Limited experience and validation in adult myositis Costly license
CHQ	Evaluates physical and psychosocial well-being	Parent or child administered	10–15 min	Computer score with proprietary algorithm	Summary score standard	No information available in juvenile DM	Good content validity, limited but good construct validity	Physical score moderately responsive	Measures important concept Psychometric properties sound Appropriate mainly for research use	Respondent burden Complicated computer scoring system Limited studies in juvenile DM

(continued)

Summary Table *(Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
Physician global damage	Overall rating of myositis disease damage based on all clinical and laboratory measures available at the time of assessment	Clinician completed	N/A	<1 min, but time to assess the patient Hand scored	On 10-cm VAS, 0 = no damage, 10 = extremely severe damage On a Likert scale, 0 = no damage, 1 = mild damage, 2 = moderate damage, 3 = severe damage, 4 = extremely severe damage	Excellent internal consistency and good interrater reliability	Good content and moderate to excellent construct validity	As expected, little responsiveness in <1 year	Measures important concept Good psychometric properties Appropriate for clinical and research use Most validated in juvenile PM/DM with limited validity in adult PM/DM	Somewhat subjective and based on the experience of the rater Reliability of serial ratings dependent on examining previous scores Not validated for IBM, and needs additional validation for adult PM/DM
MDI	Assessment of damage (persistent or permanent changes) using both VAS and present–absent scoring to assess 9 organ systems	Clinician completed	N/A	Time to complete a history and physical examination (likely 15–30 min) Hand or computer scored	For severity of damage, scores range from 0–110. For each organ system score, scored on 10-cm VAS, 0 = inactive disease, 10 = extremely severe disease activity. For the extent of damage, items are scored present or absent. Total score is 0–35 in children, 0–37 in adolescents, and 0–38 in adults. A higher score indicates more damage	Good interrater reliability Severity and extent of damage highly correlate	Good construct and criterion validity	Severity of damage score increases slowly in adult PM/DM patients, as expected Extent of damage score shows detectable mild increase	Measures important concept Sound psychometric properties Appropriate for research use in adult and juvenile PM/DM	Severity of damage scores are somewhat subjective and based on the experience of the rater For severity of damage scores, reliability of serial ratings dependent on examining previous scores Not validated for IBM
QMT	Measures amount of maximum isometric force using specialized equipment	Requires trained health care provider to conduct test	1 hour	1 hour	Values for each muscle group dependent on devices used (kg). Typically measure 8 or 12 muscle groups; total individual score for megascore	Good reliability in ALS trials and DMD; limited but good reliability in adult PM/DM	Good but very limited construct validity in IBM	Can detect changes in strength, correlated with MMT and IBMFRS SRM not available	Quantitative measure Might be sensitive to small changes in strength or in measuring mild weakness	Requires specialized training, special hardware and software; costly Patients must have at least antigravity strength to perform Very limited validation in IBM and adult PM/DM
FI-2	Assesses dynamic muscle endurance in 7 muscle groups	Observation of functional test	N/A	Takes maximum of 33 min to assess both right/left sides. Requires maximum of 21 min to assess dominant side. Takes 5 min to score by hand	Each muscle group is scored as the number of correctly performed repetitions, varying from 0–60 or 0–120. No total score	Inter- and intrarater reliability good to excellent	Good content validity Moderate construct validity	Variable; limited data from 1 therapeutic trial	Myositis-specific objective functional index that measures muscle endurance and repetition Limited validation studies in adult DM and PM Patients involved in the content validity process Measures an important concept, muscle endurance	New instrument requiring a long time to perform Further validation needed for sensitivity to change and in other myositis subgroups A training session is needed to ensure reliability Administrative burden might limit feasibility for use in clinical practice and research

(continued)

Summary Table *(Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
MAP	Assesses activity limitation and activities of daily life Contains 31 items divided into subscales	Self-reported questionnaire	5–10 min	5 min to score by hand	Subscales are scored as the median value of item responses within the subscale varying from 1 (no difficulty) to 7 (impossible) Single items are scored as the actual item response, 1–7	Moderate test–retest reliability Moderate to strong internal consistency	Good content validity Moderate construct validity	Variable; limited data from 1 therapeutic trial	Myositis-specific measure of activities of daily living and functional disability Limited validation studies in adult PM/DM Patients involved in content validity process Measures an important concept as to both difficulty and importance of activities Low administrative and patient burden support use in clinical practice and research	New measure not yet published in languages other than Swedish. Further validation needed, including sensitivity to change, construct validity, consistency of items, and performance of the tool in other myositis subgroups
IBMFRS	10-point disease-specific functional rating scale	Interviewer patient; no special training required	15 min	15 min	10 items, each 0–4 grade; add individual items for total score. 0 = several functional disabilities, 40 = no functional disability or normal function	Not available in myositis	Moderate construct validity	Very responsive in 1 therapeutic trial	Measures important elements of daily life functions that are often affected by the disease Quick, inexpensive, and easy to administer IBM-specific measure	Responses based on function prior to start of disease; subjective measurements Further validation needed
CDASI	Measures several key features of skin activity and damage in DM	Clinician completed	None	Mean 4.8 minutes for experienced dermatologists, <1 min to hand score	Scores are divided into activity and damage, with scores ranging from 0–100 for activity and 0–32 for damage. Level of disease activity can be interpreted as low, moderate, or high. Mean ± SD CDASI activity for mild disease was 11.4 ± 7.0, moderate was 25.6 ± 8.9, and severe was >39.4	Good to excellent inter- and intrarater reliability	Content validity adequate by participating dermatologists Moderate to excellent construct validity	Responsiveness strong in a group of patients with recognized change	Measures important components of skin activity and damage Psychometric properties sound Appropriate for clinical and research use Partially validated in adult DM, including amyopathic DM	Need appropriate training Does not measure every aspect of DM disease, but focuses on elements in the skin likely to be responsive in the context of therapeutic interventions Not validated in other myositis subgroups
CAT	Assesses skin disease in both activity and damage domains	Examination-based tool, administered by clinician	May take up to 15 min, depending on patient complexity and assessor's experience with DM skin disease Abbreviated tool may be faster to complete	May take up to 15 min, but scoring takes <1 min	Activity score: range 0–96, ≤1 = no activity, 7 = mild, 13 = moderate, 18 = severe, 31= very severe. Damage score: range 0–20, 0 = no damage, 1 = mild, 2 = moderate, 5 = severe or very severe	Total activity score has good internal consistency, test–retest and interrater reliability Total damage score has fair to good internal consistency, test–retest and interrater reliability Reliability of individual activity and damage items are more variable	Strong evidence for construct validity in juvenile IIM and more limited in adult DM Content validity not assessed	Responsiveness moderate to strong in children with juvenile DM with recognized change	Comprehensive assessment of relevant cutaneous lesions, including both activity and damage Partially validated in juvenile IIM and adult DM	Requires training to administer Some concerns about reliability of some items Not validated in other myositis subgroups

(continued)

Summary Table *(Cont'd)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
DSSI	Measures several key features of skin activity in DM	Clinician completed	None	2–3 min to use by experienced dermatologists, <1 min to score	Total DSSI score can range from 0–72 Compared to global physician score on a 0–10 VAS, mean ± SD DSSI was 1.28 ± 1.5 for mild global activity, 5.4 ± 4.0 for moderate global activity, and 14.9 ± 14.1 for severe global activity	Good intrarater reliability Moderate to good interrater reliability Good to excellent test–retest stability	Content validity was evaluated by a panel of experts Moderate but limited construct validity	SRM not available	Evaluates elements of skin disease activity of DM Adapted from the PASI for psoriasis Ease of use Evaluated in adult DM and amyopathic DM	BSA may not be reliable or responsive to change Does not assess skin damage Not validated in other myositis subgroups
Skindex	Measure of skin-specific QOL	Patient self-report 3 subscales: emotions, symptoms, function	5 min	Scoring involves conversion to linear scale of 100, and then taking the mean of the patient's responses in a given scale Computer scoring	Norms, as well as correlation with QOL burden in a number of different skin diseases, are available. Norms for disease severity are available	Excellent internal consistency and test–retest reliability in other skin diseases and adult DM	Moderate construct validity in DM and amyopathic DM Content and criterion validity for other skin diseases, but not available for myositis	Not available for myositis	Widely used for autoimmune, inflammatory, and noninflammatory skin diseases Correlates more highly than the SF-36 scores with the patients' reports of the condition of their skin Captures emotional component of QOL well. Limited validity for adult DM and amyopathic DM	Longer than some other skin-specific QOL measures No validation in other myositis subgroups
DLQI	Measure of skin-specific QOL	Patient self-report	2 min	Scoring takes <1 min	Score range 0–30 Interpretation can be done by cut points: 0 (score 0–1) = no effect, 1 (score 2–5) = small effect, 2 (score 6–10) = moderate effect, 3 (score 11–20) = very large effect, 4 (score 21–30) = extremely large effect	Not assessed in myositis	Moderate to low construct validity in DM and amyopathic DM Content and criterion validity not established in myositis	Not established for myositis	Widely used for autoimmune, inflammatory, and noninflammatory skin diseases Short Limited data in adult DM and amyopathic DM	Focus on disability, response distribution has ceiling effects, and dimensionality and item bias are problems Not yet studied in other myositis subgroups

* N/A = not applicable; VAS = visual analog scale; PM = polymyositis; DM = dermatomyositis; IBM = inclusion body myositis; MMT = manual muscle testing; MRC = Medical Research Council; IIM = idiopathic inflammatory myopathy; HAQ = Health Assessment Questionnaire; C-HAQ = Childhood Health Assessment Questionnaire; CMAS = Childhood Myositis Assessment Scale; MDAAT = Myositis Disease Activity Assessment Tool; MYOACT = Myositis Disease Activity Assessment VAS; MITAX = Myositis Intention to Treat Activities Index; DAS = Disease Activity Score; SF-36 = Short Form 36; HRQOL = health-related quality of life; CHQ = Childhood Health Questionnaire; MDI = Myositis Damage Index; QMT = Quantitative Muscle Testing; ALS = amyotrophic lateral sclerosis; DMD = Duchenne's muscular dystrophy; IBMFRS = Inclusion Body Myositis Functional Rating Scale; SRM = standardized response mean; FI-2 = Myositis Functional Index-2; MAP = Myositis Activities Profile; CDASI = Cutaneous Dermatomyositis Disease Area and Severity Index; CAT = Cutaneous Assessment Tool; DSSI = Dermatomyositis Skin Severity Index; PASI = Psoriasis Area and Severity Index; BSA = body surface area; QOL = quality of life; DLQI = Dermatology Life Quality Index.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S229–S239
DOI 10.1002/acr.20554

Measures of Foot Function, Foot Health, and Foot Pain

American Academy of Orthopedic Surgeons Lower Limb Outcomes Assessment: Foot and Ankle Module (AAOS-FAM), Bristol Foot Score (BFS), Revised Foot Function Index (FFI-R), Foot Health Status Questionnaire (FHSQ), Manchester Foot Pain and Disability Index (MFPDI), Podiatric Health Questionnaire (PHQ), and Rowan Foot Pain Assessment (ROFPAQ)

JODY L. RISKOWSKI,[1] THOMAS J. HAGEDORN,[2] AND MARIAN T. HANNAN[1]

INTRODUCTION

The foot is one of the most complex, yet understudied musculoskeletal systems in the body. However, with the growing interest in foot health in rheumatology and because of its pivotal role in gait and posture, researchers and clinicians have developed a number of surveys and assessments for measuring foot health and its impact on quality of life. This systematic review will focus on questionnaires and surveys for patient/participant perception of foot health and its impact on quality of life, commonly referred to as patient-reported outcome measures. The system we employed to determine the patient-reported outcome measures included in this review is provided as a flow chart (Figure 1).

AMERICAN ACADEMY OF ORTHOPEDIC SURGEONS LOWER LIMB OUTCOMES ASSESSMENT: FOOT AND ANKLE MODULE (AAOS-FAM)

Description

Purpose. To evaluate patient perception of foot health and to measure surgical outcomes (1).

Content. Questions regarding foot and ankle health from patient's perspective (1). There are 5 subscales: pain (9 questions), function (6 questions), stiffness and swelling (2 questions), giving way (3 questions), and shoe comfort (5 questions).

Number of items. 25 questions.

Response options/scale. Respondents are asked to answer on a scale of 1–5 or 1–6 with 1 being the best outcome and 5 or 6 the worst.

Recall period for items. 1 week.

Endorsements. American Academy of Orthopedic Surgeons.

Examples of use. Primarily administered to patients receiving treatment for musculoskeletal problems of the foot and ankle.

Practical Application

How to obtain. Available on the AAOS web site at URL: http://www.aaos.org/research/outcomes/outcomes_lower.asp.

Method of administration. Self-administered.

Scoring. Scoring spreadsheet and instructions are available with the assessment. Scores are standardized to a percentage (0–100) score and then transformed on normative scale. Scoring is automated on available worksheet.

Score interpretation. A lower normative score indicates worse foot health relative to the population (2). Scores range from 0–100 for each subscale and can be placed on a normative scale from −26 to 56 based on the general population (1,2). The mean ± SD population score for the global foot and ankle module is 93.19 ± 12.33 (n = 1,755) (2).

Respondent burden. Not reported.

Administrative burden. Training consists of self-study of the scoring documentation (see URL: http://www.aaos.org/research/outcomes/outcomes_lower.asp).

[1] Jody L. Riskowski, PhD, Marian T. Hannan, DSc, MPH: Hebrew SeniorLife, and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts; [2] Thomas J. Hagedorn, BS: Hebrew SeniorLife, Boston, Massachusetts.

Address correspondence to Marian T. Hannan, DSc, MPH, Co-Director of Musculoskeletal Research, Institute for Aging Research, Hebrew SeniorLife, 1200 Centre Street, Boston, MA 02131-1097. E-mail: hannan@hsl.harvard.edu.

Submitted for publication February 17, 2011; accepted in revised form July 5, 2011.

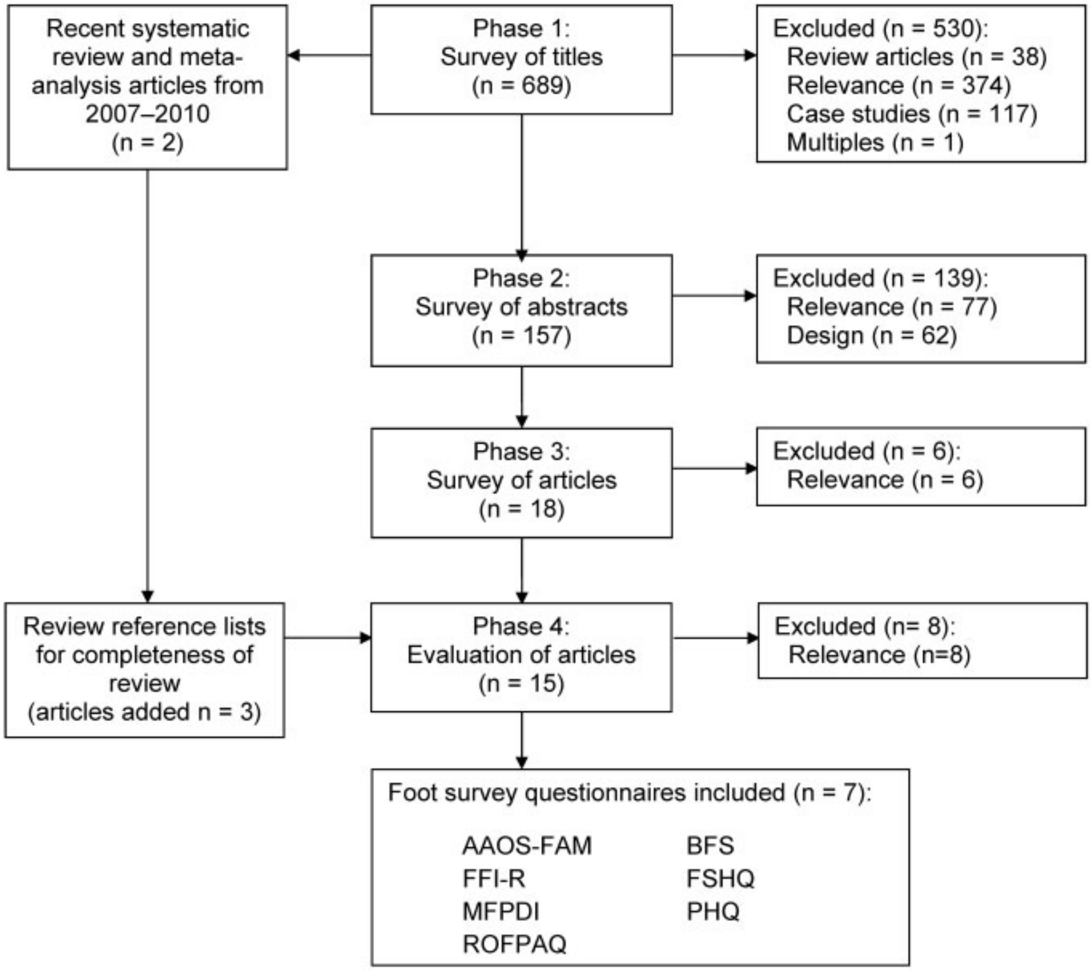

Recent systematic review and meta-analysis articles from 2007–2010 (n = 2)
Phase 1: Survey of titles (n = 689)
Excluded (n = 530): Review articles (n = 38) Relevance (n = 374) Case studies (n = 117) Multiples (n = 1)
Phase 2: Survey of abstracts (n = 157)
Excluded (n = 139): Relevance (n = 77) Design (n = 62)
Phase 3: Survey of articles (n = 18)
Excluded (n = 6): Relevance (n = 6)
Review reference lists for completeness of review (articles added n = 3)
Phase 4: Evaluation of articles (n = 15)
Excluded (n= 8): Relevance (n=8)
Foot survey questionnaires included (n = 7):
AAOS-FAM
BFS
FFI-R
FSHQ
MFPDI
PHQ
ROFPAQ

Figure 1. Identification of studies for inclusion in the review. AAOS-FAM = American Academy of Orthopedic Surgeons Lower Limb Outcomes Assessment: Foot and Ankle Module; BFS = Bristol Foot Score; FFI-R = Revised Foot Function Index; FHSQ = Foot Health Status Questionnaire; MFPDI = Manchester Foot Pain and Disability Index; PHQ = Podiatric Health Questionnaire; ROFPAQ = Rowan Foot Pain Assessment.

Translations/adaptations. Full assessment is split into several submodules that include questionnaires evaluating the lower-extremity core, foot and ankle, hip and knee, sports-related injuries, and common knee problems (1).

Psychometric Information

Method of development. Content was developed and refined with input from clinician focus groups (1).

Acceptability. Not reported.

Reliability. Internal Cronbach's alpha of 0.91, 0.83, 0.61, and 0.88 was reported for the pain, function, stiffness, and giving way subscales, respectively, and 0.93 for the entire foot and ankle module. With the exception of the stiffness subscale, these indicate generally good internal reliability. The module had a test–retest reliability measured internally as 0.79, and subscale test–retest reliability of 0.87, 0.81, 0.99, and 0.81 for the pain, function, stiffness, and giving way subscales, respectively (1). In an independent study of reliability, Hunsaker et al (2) reported Cronbach's alpha of 0.81–0.96 for all lower-extremity core (foot and ankle, hip and knee, sports-related injuries, and common knee problems, respectively) without noting the individual subscale values. Their reported test–retest reliability was 0.79 for the foot and ankle module (2).

Validity. The questionnaire was validated by comparison with clinical assessments performed by a trained physician, and correlations between the questionnaire and physician scores of pain (r = 0.49) and function (r = 0.43) were observed. Patient responses were also seen to be strongly correlated with Short Form 36 (SF-36) scores (r = 0.65) and assessment of the lower-extremity core (r = 0.89) (1).

Ability to detect change. No data have been reported on the ability of the global foot and ankle modules to detect change; however, overall lower-extremity scores were shown to correlate (r = 0.54) with changes in physician-assessed function scores indicating responsiveness to change (1).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. AAOS-FAM is one of the few foot patient-reported outcome measures that have internal and external reliability measures.

Caveats and cautions. This questionnaire does not evaluate the impact of foot health with regard to its impact on the participant's psychological state, social activities, or self-esteem, all of which may influence quality of life and patient satisfaction (3).

Clinical usability. This survey was designed for orthopedists and health care professionals to validate and compare results and clinical outcomes across studies (4). As the AAOS-FAM is clinical in nature, few questions address quality of life; however, by combining the AAOS-FAM with the SF-36, the 2 instruments can be a means for evaluating foot health–related quality of life (1). Further, the AAOS-FAM, similar to several other foot-related patient-reported outcome measures, lacks an independent review of the validity and lacks information regarding the minimum detectable difference and minimum clinically important difference, which limits its clinical usability.

Research usability. Most studies that have used the AAOS-FAM have focused on outcomes assessment concerning treatment of a particular condition (e.g., clubfoot [5]) or of surgical method (e.g., Ilizarov method for tibial nonunions [6]). However, because it was designed to measure clinical assessments, its usability for assessing population-level or community-based foot and ankle health appears limited.

BRISTOL FOOT SCORE (BFS)

Description

Purpose. To assess the patient's perception of the impact of foot problems on everyday life (7).

Content. Questions relating to foot pain and concern, footwear and general foot health, and mobility. There are 3 subscales: foot concern and pain (7 questions), footwear and general foot health (4 questions), and mobility (3 questions) (7). Fourteen of the 15 questions are scored; the final question is a statement of general health, which does not add into the BFS.

Number of items. 15 questions.

Response options/scale. Each response option is assigned a score of 1 (best possible situation) to 3–6 (worst possible situation, number dependent on number of response options available) for each BFS survey question (7).

Recall period for items. 2 weeks.

Endorsements. None.

Examples of use. Target population is podiatric patients, and it has been used to study effects of nail fungus treatment (8) and foot surgery (9).

Practical Application

How to obtain. Available in the original article (7).

Method of administration. Self-administered.

Scoring. Scores for each question are summed per a provided scoring guide. Scores range from 15 (best possible situation) to 73 (worst possible situation). Within the subscales, foot concern and pain scores range from 7–36, footwear and general foot health scores range from 4–20, and mobility scores range from 3–12.

Score interpretation. Lower scores indicate that the patient perceives fewer foot problems.

Respondent burden. 3–5 minutes to complete (7).

Administrative burden. Training consists of self-study of the scoring documentation (7).

Translations/adaptations. English only.

Psychometric Information

Method of development. Topic-guided interviews with podiatric patients (7).

Acceptability. Not reported.

Reliability. The survey developers noted a combined Cronbach's alpha of 0.90 for the BFS, and the Cronbach's alpha for the individual subscales was not reported. Test–retest values from 36 patients over a 2-week wait-list period were −0.83; test–retest reliability of the individual subscales is unknown (7).

Validity. Content validity was evaluated by comparing the BFS with a clinical evaluation using the United Bristol Healthcare National Health Service Trust standard content validity with the Chiropody Assessment Criteria Score in a group of 54 podiatric patients (41 women and 13 men). There was a negligible, nonsignificant correlation between these scores with an $r = 0.14$, which suggests that these measures reflect different outcomes (5).

Ability to detect change. Barnett et al showed a BFS pre-post change of 1.2 ± 7.1 for the 54 patients after 2 weeks of routine care. In 49 patients (25 women and 24 men), there was an 18.7 ± 12.3 point pre-post change in the 6 weeks following nail surgery in their BFS ($P = 0.01$) (7). However, there are no independent studies determining the minimum detectable or minimum clinically important difference.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BFS was developed based on patients' perspectives of foot health and ailments, which provides it better content validity for assessing complaints.

Caveats and cautions. Psychometric evaluation for the BFS is limited, and there is no independent assessment of its psychometric properties. The 3 subdomains (i.e., foot concern and pain, footwear and general foot health, and mobility) do not show construct validity against other foot questionnaires or against a clinical assessment.

Clinical usability. The BFS was developed with focus groups, but without an independent study of its psychometric properties and without known values of the minimum detectable or minimum clinically important difference. The clinical utility of the BFS may be limited.

Research usability. Campbell (10) suggests that because the BFS was developed in a clinical setting, it is not as useful for monitoring the change in foot health in populations with a low risk of foot ailments.

REVISED FOOT FUNCTION INDEX (FFI-R)

Description

Purpose. To assess foot-related health and quality of life.

Content. Questions to evaluate overall foot function, foot health, and quality of life. The FFI-R has 4 subscales: pain and stiffness (19 questions), social and emotional outcomes (19 questions), disability (20 questions), and activity limitation (10 questions). The FFI has 3 subscales: pain (9 questions), disability (9 questions), and activity limitation (5 questions).

Number of items. Long-form FFI-R consists of 68 questions. Shorter form has 34 questions that only assess foot function, and it is not intended for analysis of subscales (11). The original FFI consists of 23 items on 3 subscales (12).

Response options/scale. FFI-R respondents answer on a Likert scale of 1–5. Some items also contain a sixth possible response indicating that it is not applicable to the respondent (11). FFI is scored on a visual analog scale between verbal anchors representing extremes (12).

Recall period for items. 1 week.

Endorsements. None.

Examples of use. Patients with rheumatoid arthritis (11), but it has also been used to assess orthotics outcomes (13).

Practical Application

How to obtain. Available in original publication (11).

Method of administration. Self-administered.

Scoring. If the 68-question FFI-R is administered, an index is calculated by summing responses and dividing by the maximum possible score on each subscale to obtain separate percentage scores for each. The 34-question FFI-R is used to obtain an overall score of foot function (11). On the FFI, visual scales are divided into 10 equal segments and the respondents' mark classified as a number between 0 and 9. Scores are then summed on subscales, and evaluated as a percentage of the highest possible score (12).

Score interpretation. Range of 0–100% on each subscale, plus an overall percentage score. Higher scores indicate worsening foot health and poorer foot-related quality of life on both the FFI-R and FFI (11,12).

Respondent burden. Less than 30 minutes to complete (11).

Administrative burden. Self-study of the scoring documentation (11).

Translations/adaptations. FFI-R has 2 versions (long form and short form); previous version is FFI (12).

Psychometric Information

Method of development. Adapted from information obtained from previous survey, patient focus groups, and foot specialists (11).

Acceptability. The questionnaire is written for an eighth-grade reading level.

Reliability. The survey developers noted the FFI-R test–retest person reliability was 0.96 and the item reliability was 0.93. The developers also reported Cronbach's alpha of 0.93, 0.86, 0.93, and 0.88 for the pain, psychosocial, disability, and functional limitation subscales, respectively, indicating high internal reliability (11). The FFI survey developers reported the FFI as having a high test–retest reliability, with an intraclass correlation coefficient (ICC) of 0.87 for the full questionnaire. Subscale ICCs were 0.69, 0.81, and 0.84 for the pain, disability, and activity limitation subscales, respectively. Budiman-Mak et al reported the FFI Cronbach's alpha as 0.96 for the full questionnaire, with subscale alpha of 0.73, 0.93, and 0.95 for the activity limitation, disability, and pain subscales, respectively, indicating high internal reliability (12).

Validity. FFI-R results were compared to a 50-foot walking time (11,12). Significant correlation was observed between walk times and the FFI-R score ($r = 0.31$, $P = 0.018$). The construct validity was also supported by the correspondence of items considered to indicate low severity of problems being associated with lower scores (indicating better foot health and function) (11). Factor analysis of the FFI showed overall construct validity, with all but 2 items weighing into a single factor. Analysis with varimax rotation also showed subscale validity, with all pain and disability items separating into 2 factors, and activity limitation items dividing between 2 additional factors. Content validity was gauged by correlation with 50-foot walk times and counts of painful joints. The FFI had a moderate overall correlation of 0.48 and 0.53 when compared to walk times and painful joint counts, respectively (12).

Ability to detect change. Minimum detectable difference and minimum clinically important difference have not been reported for the FFI-R. The pain and activity limitation subscales of the FFI have been correlated to changes in the number of painful joints over 6 months ($r = 0.47$, $P = 0.002$ and $r = 0.34$, $P = 0.03$, respectively). There was no significant relationship observed between the disability subscale and the number of painful joints ($r = 0.11$, $P = 0.51$) (12). In an independent study examining treatment of plantar fasciitis in 175 patients, Landorf and Radford (14) found the minimally important difference on the FFI was −0.5 points for activity limitation, −12.3 points for pain, and −6.7 points for disability, with a total FFI change of −6.5. Negative scores denoted improved foot-related health.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The FFI-R provides both a short and long form, which provides the researcher an option of the level of detail necessary.

Caveats and cautions. FFI-R is a questionnaire based on the original FFI, seeking to address criticisms relating to the original index's basis, administrative issues, validity, and psychometric properties (11,15). Though based on the FFI, the FFI-R is a notably different survey in length, construction, and content. While the FFI-R is the newer survey, many researchers continue to use the older, more established FFI. However, because the FFI and FFI-R are different, it is difficult to compare results between these surveys.

Clinical usability. The FFI-R was developed through patient and focus groups, but its validity, reliability, and sensitivity to change have not been independently evaluated. The FFI-R, similar to several other foot-related patient-reported outcome measures, lacks an independent review of the psychometric properties and lacks information regarding the minimum detectable difference and minimum clinically important difference, which limits its clinical usability.

Research usability. The FFI-R was developed from the original FFI and a literature review, as well as focus groups with foot specialists, interviews with foot specialists and podiatric patients, and results from patient surveys (11). As a result, the FFI-R is noted to be a well-developed measure of foot health–related quality of life (16); however, because it is also a newer survey, there are fewer independent studies evaluating its utility.

FOOT HEALTH STATUS QUESTIONNAIRE (FHSQ)

Description

Purpose. To measure foot health related to quality of life (17,18).

Content. Questions regarding foot health and its impact on quality of life. There are 4 subscales: foot pain (4 questions), foot function (4 questions), footwear (3 questions), and general foot health (2 questions).

Number of items. 13 questions.

Response options/scale. For the subscales of pain, function, and general foot health, a 5-point Likert scale of no problems, pain, or limitations to severe problems, pain, or limitations. Responses to footwear questions are on a 5-point bipolar Likert scale from strongly disagree to strongly agree for statements regarding shoe fit, discomfort wearing shoes, and shoewear available.

Recall period for items. 1 week.

Endorsements. None.

Examples of use. Used to assess the effects of footwear (19) and orthotic interventions (20,21), and foot health in the community (22), as well as in various podiatric clinical populations (23–25).

Practical Application

How to obtain. Survey and scoring program are available through the FHSQ web site at URL: http://fhsq.homestead.com/index.html. Its current price is AUS $150.

Method of administration. Self-administered.

Scoring. Dedicated FHSQ program scores questionnaires. When fewer than 50% of the responses for any one scale are missing, the missing responses are assigned with the average value of the completed questions for that scale (17).

Score interpretation. Subscale scores are reported as 0 (poorest state of foot health) to 100 (optimal foot health). Higher scores reflect better foot health and quality of life (17,18).

Respondent burden. Less than 10 minutes to complete.

Administrative burden. Not reported.

Translations/adaptations. Original in English (17,18), with translated versions in Brazilian Portuguese (26) and Spanish (Valencian culture) (27).

Psychometric Information

Method of development. Content was developed with input from focus groups of podiatric surgeons.

Acceptability. Not reported.

Reliability. The survey developers reported the FHSQ Cronbach's alpha for subscales was 0.85 (footwear), 0.86 (foot function), 0.88 (general foot health), and 0.88 (foot pain) in a sample of 111 podiatric patients (18) and 0.89–0.95 (individual alpha for each subscale not provided) (17). These alphas were between the accepted 0.7–0.9 range (28). The survey developers noted the intraclass correlations were 0.74 (footwear), 0.78 (general foot health), 0.86 (foot pain), and 0.92 (foot function) for the test–retest reliability of 72 patients who completed the survey before and after a week of routine care, noting a high reliability (18).

Validity. The survey developers assessed validity with 111 podiatric patients. The root mean standard error of approximation was 0.08, which suggests a moderate fit of the FHSQ to measure foot health related to quality of life (29). The goodness-of-fit index, an absolute index of fit, was 0.90, while the comparative fit index (CFI), a relative measure of fit, was 0.96 (17). The CFI depends on the average size of the correlations in the data, so a high value suggests a high correlation between variables. The CFI was above the recommended 0.95 cutoff (30), suggesting high validity.

Ability to detect change. In an independent study examining treatment of plantar fasciitis in 175 patients, Landorf and Radford (14) found the minimally important difference for pain was 14 points (i.e., pain scores increased by 14 points), for function was 7 points, and general foot health was 9 points to denote improved foot-related health. An independent study also evaluated the clinically relevant responsiveness of the FHSQ foot function subscale in 784 ethnically diverse older adults (31). In this study, the FHSQ foot function subscale scores differed between 3 groups of participants. Participants in one group with minor foot pathology (e.g., hyperkeratosis and nail pathology) had a mean FHSQ foot function subscale score of 88.8. Participants who had a morphologic disorder (e.g., hammertoes) had a mean FHSQ foot function subscale score of 77.9. Participants in a third group with acute disease (e.g., plantar fasciitis) had an average FHSQ foot function subscale score of 53.9. The decrements of FHSQ scores associated with an increasing number of foot disorders in this study ranged between 10 and 20 points, similar to the differences reported earlier. These results suggest that the changes in foot function FHSQ subscores are clinically relevant to poorer foot function as a result of an increasing number of foot disorders (18,31).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The 4 subscales are representative of health and health impact on quality of life and disability (32,33). Moreover, the FHSQ has more psychometric data available compared to others (16) and is used within a number of research settings, despite its cost.

Caveats and cautions. Trevethan argues that better psychometric analyses would allow for some questions to be removed and could reduce the participant burden (15). Further, this questionnaire does not evaluate the impact of foot health with regard to its impact on the participant's psychological state, social activities, or self-esteem, all of which may influence quality of life and patient satisfaction (3).

Clinical usability. With known values of the minimal important difference, as well as many of the psychometric properties, the FHSQ is frequently used in clinical settings.

Research usability. With high validity and an independent study assessing minimal important differences, this foot-related patient-reported outcome measure has well-detailed psychometric properties and is one of the most common foot surveys.

MANCHESTER FOOT PAIN AND DISABILITY INDEX (MFPDI)

Description

Purpose. To measure disabling foot pain in the general population (34).

Content. Questions of foot health as they relate to foot pain, functional limitations, and self/body image. The original survey has 3 subscales: functional limitation (10 questions), pain intensity (5 questions), and perception of one's appearance as a result of foot problems (2 questions) (34). Menz et al performed an independent factor analysis using a sample of 301 older adults in Australia, which showed 4 subscales: functional limitation (7 questions), activity restrictions (2 questions), pain (6 questions), and concern over foot appearance (2 questions) (35). The Manchester-Oxford Foot Questionnaire (MOXFQ) showed 3 subscales from a factor analysis: walking/standing domain (7 items), pain (5 questions), and social interactions (4 items) (36). The factor analysis by Cook et al noted 2 subscales: foot and ankle function (9 questions) and pain and appearance (7 questions) (37).

Number of items. 17 questions in the original (34) or 16 questions after a separate item response theory analysis (37). The MOXFQ also has 16 questions (36).

Response options/scale. Responses have 3 levels of severity (never, sometimes, always), which are transformed into numerical scores (and summed within each subscale).

Recall period for items. 1 month.

Endorsements. None.

Examples of use. Used as a general population survey of adults and older adults (35) to evaluate disabling foot pain (34,35,38) or hallux valgus surgery (36).

Practical Application

How to obtain. Available in the original publication (34).

Method of administration. Self-administered or interview (34).

Scoring. Items are summed per scoring guide of version used (34–37). Original publication assigns the severity level values of 1–3, corresponding to increasing severity (34). Subsequent publications have also evaluated an overall score expressed as the sum of each subscale score or as a percentage of the total possible outcome (35).

Score interpretation. The range varies depending on the scoring technique used, and original survey used a 0–2 scale, yielding a score range of 0–34 (34). Cook et al and Waxman et al used a 1–3 scoring for range of 17–51 (37,39). Higher scores correspond to more severe foot pain and disability (34).

Respondent burden. Not reported.

Administrative burden. Self-study of the scoring documentation (34–37).

Translations/adaptations. Original is in English; Greek (40), Italian (41), and Brazilian Portuguese (42) versions have also been validated. The MOXFQ was developed from the MFPDI to assess hallux valgus corrective surgery (36). Cook et al performed a graded response item response theory analysis to reduce the MFPDI by 1 less question (37).

Psychometric Information

Method of development. Open-ended interviews with 32 patients who visited a foot clinic (34).

Acceptability. Not reported.

Reliability. Garrow et al (34) reported a Cronbach's alpha of 0.99 (34), whereas an independent study noted it as 0.89 (35), indicating high reliability. Both research groups stated the questionnaire has high consistency (no statistics provided) with self-report of injury during separate patient interviews in younger and older populations (34,35). In the MOXFQ survey, Dawson et al reported Cronbach's alpha coefficients of 0.73, 0.86, and 0.92 for the social interaction, pain, and walking/standing subscales, respectively, when evaluating 100 hallux valgus surgery patients (36). These Cronbach's alpha coefficients were between the accepted 0.7–0.9 range (28).

Validity. Content of the survey was generated with patient interviews, and the construct validated through the comparison of responses from groups with known systematic differences in foot conditions. The criteria of the MFPDI were also compared to similar items in the ambulation subscale of the Function Limitation Profile Questionnaire. This comparison showed that items with similar wording had a Cohen's kappa of 0.48 and 0.50, and a much lower kappa (0.17) for differently worded items (34). Cohen's kappa is a measure of agreement with higher values indicating better agreement, with moderate agreement ranging from 0.4–0.6 and slight agreement <0.2 (43). The functional limitation and activity restriction subscales have been shown to be significantly correlated with the Short Form 36 (SF-36) mental ($r = 0.20$, $P = 0.04$) and general (r = 0.21, P = 0.03) health subscales (35). The Dawson et al study of 100 hallux valgus surgery patients also assessed the MOXFQ validity (36). MOXFQ walking/standing subscale was strongly associated ($P < 0.001$) with the SF-36 physical functioning (Spearman's correlation r = 0.68), role physical (r = 0.58), and pain (r = 0.54) domains, and with the SF-36 physical component summary score (r = 0.63). The MOXFQ was strongly associated ($P < 0.001$) with the SF-36 pain subscale (r = 0.53).

Ability to detect change. The original MFPDI does not have reported sensitivity, responsiveness, or minimal important difference data. The MOXFQ assessment of corrective hallux valgus treatment does provide data regarding the subscale minimally important differences. Dawson et al noted the minimum clinically important differences were 12.8 points (effect size 0.4), 4.6 points (effect size 0.2), and 20.3 points (effect size 0.8) for the walking/standing, pain, and social interaction subscales, respectively. In evaluating pain transition, receiver operating characteristic curves provided cut points for the MOXFQ. The suggested cut points were 14 points for the walking/standing scale and 25 points for both the pain and social interaction scales to indicate a minimally important amount of change.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. MFPDI measures foot pain and functional limitations from multiple perspectives and with multiple questions, which provides an appropriate means for reducing measurement error (44).

Caveats and cautions. The MFPDI provides only 2 questions for addressing footwear, and there are no questions regarding self/body image. Because footwear can affect self/body image (45), this questionnaire may not capture the effects of footwear or footwear interventions from the patient's perspective.

Clinical usability. There are several different assessment models and adaptations of the MFPDI developed. However, within these surveys and scoring methods, only the MOXFQ has minimally important differences noted in populations with hallux valgus. The other adaptations from the MFPDI should be further independently evaluated for their minimum detectable difference and minimum clinically important difference to improve their clinical utility.

Research usability. Menz et al noted there were 4 subscales instead of 3 for their population of older adults (35). In an independent analysis of the 3 assessment models (3 domains in the Garrow et al original study [34], 4 subscales in the Menz et al study of older adults [35], and 2 domains of the Cook et al study [37]), the Garrow et al study performed better (lower root mean square error of approximation [0.065], higher comparative fit index [0.949], and higher normed fit index [0.943]) than the other 2 studies in a survey of adults over age 50 (46). Therefore, the correct scoring model should be evaluated relevant to the population studied.

PODIATRIC HEALTH QUESTIONNAIRE (PHQ)

Description

Purpose. To measure foot-related health in podiatric patient populations (47).

Content. Questions related to walking, foot health, foot pain, worry about feet, and impact of the foot on quality of life. Includes 7 subscales: walking, foot hygiene, nail care, foot pain, worry about feet, and impact on quality of life, with one question each and separate visual analog scale (VAS) for current foot status.

Number of items. 6 questions and 1 VAS, for a total of 7 items.

Response options/scale. Each dimension has 1 question related to it with 3 severity levels (no problems, some problems, and severe problems). 20-cm VAS delineated from 0–100.

Recall period for items. 1 day.

Endorsements. None.

Examples of use. PHQ has been used in podiatric patient populations with various foot ailments and systemic diseases, such as rheumatoid arthritis and diabetes (47,48).

Practical Application

How to obtain. Available in the original article (47).

Method of administration. Self-administered.

Scoring. 6 dimensions are summed per scoring guide to generate a single score ranging from 6–18.

Score interpretation. Higher scores indicate more severe problems, and a higher VAS score indicates better foot health. Scoring is categorical, based on the level of severity (level 1 = no problems to level 3 = severe problems). The VAS is delineated from 0 (worst possible foot health) to 100 (best possible foot health) for the response item "How are your feet today?" (47).

Respondent burden. Not reported.

Administrative burden. Training of the podiatric staff for the PHQ and clinical podiatric assessment is 2 hours (47).

Translations/adaptations. English only.

Psychometric Information

Method of development. Consultation of podiatric managers and podiatric clinicians (47).

Acceptability. Not reported.

Reliability. Unknown.

Validity. The survey developers validated the PHQ against the generic health status assessment of the EuroQol 5-Domain instrument (EQ-5D) and an objective clinical assessment in which a podiatrist objectively scored the patient's foot health from 1 (no foot problems) to 5 (severe foot problems) (47). Comparing the PHQ to the clinical podiatric assessment, the Goodman-Kruskal lambda for the 2,038 patients for each dimension was: walking 0.15, hygiene −0.09, nail care −0.24, foot pain 0.41, worry/concern for feet 0.30, and impact on quality of life 0.31. The PHQ was noted to be more robust in detecting foot-related health than the EQ-5D when it was compared to the clinical podiatric assessment (the subscale Goodman-

Table 1. Content of patient-reported foot health questionnaires*

	Foot pain	Foot health	Foot function	Functional limitation/disability	Self-perception/ body image	Psychological	Social	Orthotics/ shoewear
AAOS-FAM	Yes	Yes	Yes	Yes	–	–	–	Yes
BFS	Yes	Yes	–	Yes	Yes	Yes	Yes	Yes
FFI-R	Yes	–	–	Yes	–	Yes	Yes	–
FHSQ	Yes	Yes	Yes	Yes	–	–	–	Yes
MFPDI	Yes	–	–	Yes	Yes	–	–	–
PHQ	Yes	Yes	–	Yes	–	Yes	–	–
ROFPAQ	Yes	–	–	Yes	–	Yes	Yes	–

* AAOS-FAM = American Academy of Orthopedic Surgeons Lower Limb Outcomes Assessment: Foot and Ankle Module; BFS = Bristol Foot Score; FFI-R = Revised Foot Function Index; FHSQ = Foot Health Status Questionnaire; MFPDI = Manchester Foot Pain and Disability Index; PHQ = Podiatric Health Questionnaire; ROFPAQ = Rowan Foot Pain Assessment.

Kruskal lambda ranged from 0.13–0.02) (47). Goodman-Kruskal lambda is a measure of the proportional ability of predicting the outcome for 1 categorical variable based on a second categorical variable. For construct validity, the PHQ subscales were correlated to the EQ-5D components ranging from 0.58–0.14 using Kendal correlation coefficients, and the PHQ_{vas} and EQ-$5D_{vas}$ had a 0.40 Kendal correlation coefficient (47). These values suggest a low to moderate correlation, suggesting that the PHQ and EQ-5D detect different aspects of health.

Ability to detect change. In an independent study, Farndon et al used the PHQ to determine changes in foot status over a 2-week period after a podiatric intervention of 1,047 patients in 8 podiatric clinics (48). In 2 weeks, they noted a significant ($P < 0.001$) change in the PHQ dimension scores and the PHQ_{vas} for their patients. The PHQ of the 6 dimensions decreased by 0.5 (95% confidence interval [95% CI] 0.4–0.7). The PHQ_{vas} decreased by 0.7 (95% CI 0.6–0.9) using the PHQ_{vas} on a 0–10 scale (no pain to worse pain). While they initially used a clinical assessment to validate their PQH and PHQ_{vas} scores, in the followup PQH assessment, there was no followup clinical assessment to assess the validity of the change in scores. Therefore, the minimum detectable difference and minimum clinically important difference are both unknown.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. In terms of the number of survey questions, the PHQ is one of the shortest foot-related patient-reported outcome measures, which can limit the participant burden.

Caveats and cautions. The PHQ is a 1 question per domain measurement of foot health. This allows for patients and survey participants to quickly take the questionnaire; however, this may also increase measurement error because there is no means of ensuring the question was understood or was a representative answer of the impact of foot health on the patient's quality of life (44).

Clinical usability. Without known minimum detectable difference and minimum clinically important difference, the clinical utility of this survey is limited. Further, there are no questions regarding foot function, orthotics, and shoewear, all of which are important features of podiatric treatment and evaluation.

Research usability. Perhaps due to the sparseness of this survey with regard to the number and type of questions, this survey is not commonly used in research settings.

ROWAN FOOT PAIN ASSESSMENT (ROFPAQ)

Description

Purpose. To evaluate chronic foot pain (49).

Content. Addresses the 3 pain dimensions: sensory, affective (motivational), and cognitive (49,50). 3 subscales: sensory (16 questions), affective (10 questions), and cognitive (10 questions), with 3 additional questions used as indicators of understanding.

Number of items. 39 questions.

Response options/scale. Each question has a Likert scale from 1 (no foot pain or foot pain does not affect patient) to 5 (extreme foot pain or foot pain significantly affects patient). The subscale questions (i.e., sensory, affective, and cognitive) are distributed throughout the questionnaire in lieu of being grouped by domain, and they should be scored within each subscale (49). The 3 comprehension questions should be assessed to see if they are similar.

Recall period for items. Unspecified.

Endorsements. None.

Examples of use. Podiatric patients with chronic foot pain.

Practical Application

How to obtain. Available in the appendix of the original article (49).

Method of administration. Self-administered.

Scoring. Scores within each subdomain are summed, with the sensory domain score ranging from 16–80, and the affective and cognitive domains ranging from 10–50.

Score interpretation. Higher scores suggest that foot pain has a greater effect on the patient's pain domains and is less ideal for the patient. The 3 comprehension questions should have a 90% agreement; if comprehension scores are less than 90%, the survey administrator should have the patient retake the survey or verbally clarify the statements since it may indicate either patient carelessness or question misunderstanding.

Respondent burden. Mean completion time is 9 minutes (range 2–20 minutes) (49).

Administrative burden. Self-study of the scoring documentation (49).

Translations/adaptations. English only.

Psychometric Information

Method of development. Data from 6 focus groups and 2 semistructured interviews used to guide development (49).

Acceptability. Reported Flesch reading ease score of 74.8, which is slightly better than average readability (49).

Reliability. Thirty-nine participants (26 women and 13 men) with foot pain for more than 1 year took the ROFPAQ survey to assess reliability and validity measures. The survey developer noted the internal consistency scores were 0.90 (sensory), 0.81 (affective), and 0.87 (cognitive), between the accepted values of 0.7 and 0.9 (49). The Spearman's test–retest reliability coefficients were 0.88 (sensory), 0.93 (affective), and 0.82 (cognitive) when participants took the ROFPAQ twice, 24 hours apart, indicating high reliability.

Validity. Validity, the ability of the survey to detect chronic foot pain over other types of pain, was supported in that the survey distinguishes the effects of chronic foot pain over headache pain. To measure convergent validity, the ROFPAQ was compared to the Foot Function Index (FFI) pain subscale (12); the Spearman's correlation coefficients between these scales were 0.88 (sensory), 0.69 (affective), and 0.70 (cognitive). As the subdomains of the ROFPAQ were correlated to the FFI pain measure, the author states that this suggests that the ROFPAQ measures more than the sensory domain of pain (49). No independent studies have examined the validity of the ROFPAQ.

Ability to detect change. Unknown.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The ROFPAQ was designed and validated to assess the 3 domains of foot pain, and it does evaluate pain from multiple perspectives (sensory, affective [motivational], and cognitive).

Caveats and cautions. Since this survey was only designed to assess foot pain, it does not measure the commonly associated features of foot pain (e.g., foot function, foot health, and shoewear).

Clinical usability. The ROFPAQ was designed to measure the 3 dimensions of chronic foot pain; as a result, this assessment does not model foot health on quality of life as well as other questionnaires. Therefore, it is best suited for assessing treatment modalities in podiatric clinical populations as opposed to community-based studies of foot health.

Research usability. The ROFPAQ does not have an independent study of its psychometric properties, and the survey is not commonly used, which limits the ability to evaluate results across research and clinical populations. Further, because the survey only measures foot pain without regard to other commonly associated features of foot pain (e.g., foot function, foot health, or shoewear), it suggests that including a separate survey or set of questions regarding these aspects may be necessary to fully evaluate the role of foot pain on the participant's life.

DISCUSSION

This review has described several of the instruments used to measure foot-related patient-reported outcome measures in adults. Table 1 lists the content comparisons of these foot health questionnaires. Currently, the area of foot health and foot function is garnering greater attention in the rheumatology community. Thus, there is a great need for valid and reliable instruments and surveys to measure foot health. However, many of the foot-related patient-reported outcome measures have limited evidence regarding their validity and responsiveness to change, limiting their use in clinical intervention and population studies. It is important to note that this review is limited to instruments primarily used in adults, and further work is needed to include pediatric measures. Future work should evaluate the psychometric properties and clinical utility of these foot-related patient-reported outcome measures.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Johanson NA, Liang MH, Daltroy L, Rudicel S, Richmond J. American Academy of Orthopaedic Surgeons lower limb outcomes assessment instruments: reliability, validity, and sensitivity to change. J Bone Joint Surg Am 2004;86-A:902–9.
2. Hunsaker FG, Cioffi DA, Amadio PC, Wright JG, Caughlin B. The American academy of orthopaedic surgeons outcomes instruments: normative values from the general population. J Bone Joint Surg Am 2002;84-A:208–15.
3. Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med 1999; 48:1507–15.
4. Keller RB. AAOS/council of musculoskeletal specialty societies' outcomes instruments. Curr Opin Orthop 1996;7:55–60.
5. Wallander H, Larsson S, Bjonness T, Hansson G. Patient-reported outcome at 62 to 67 years of age in 83 patients treated for congenital clubfoot. J Bone Joint Surg Br 2009;91:1316–21.
6. Brinker MR, O'Connor DP. Outcomes of tibial nonunion in older adults following treatment using the Ilizarov method. J Orthop Trauma 2007; 21:634–42.
7. Barnett S, Campbell R, Harvey I. The Bristol Foot Score: developing a patient-based foot-health measure. J Am Podiatr Med Assoc 2005;95: 264–72.
8. Malay DS, Yi S, Borowsky P, Downey MS, Mlodzienski AJ. Efficacy of debridement alone versus debridement combined with topical antifungal nail lacquer for the treatment of pedal onychomycosis: a randomized, controlled trial. J Foot Ankle Surg 2009;48:294–308.
9. Sugathan HK, Sherlock DA. A modified Jones procedure for managing clawing of lesser toes in pes cavus: long-term follow-up in 8 patients. J Foot Ankle Surg 2009;48:637–41.
10. Campbell JA. Characteristics of the foot health of "low risk" older people: a principal components analysis of foot health measures. Foot 2006;16:44–50.
11. Budiman-Mak E, Conrad K, Stuck R, Matters M. Theoretical model and

Rasch analysis to develop a revised Foot Function Index. Foot Ankle Int 2006;27:519–27.

12. Budiman-Mak E, Conrad KJ, Roach KE. The Foot Function Index: a measure of foot pain and disability. J Clin Epidemiol 1991;44:561–70.
13. Rao S, Baumhauer JF, Tome J, Nawoczenski DA. Orthoses alter in vivo segmental foot kinematics during walking in patients with midfoot arthritis. Arch Phys Med Rehabil 2010;91:608–14.
14. Landorf KB, Radford JA. Minimal important difference: values for the Foot Health Status Questionnaire, Foot Function Index and Visual Analogue Scale. Foot 2008;18:15–9.
15. Trevethan R. Evaluation of two self-referent foot health instruments. Foot (Edinb) 2010;20:101–8.
16. Walmsley S, Williams A, Ravey M, Graham A. The rheumatoid foot: a systematic literature review of patient-reported outcome measures. J Foot Ankle Res 2010;3:12.
17. Bennett PJ, Patterson C. The foot health status questionnaire (FHSQ): a new instrument for measuring outcomes of foot care. Australasian J Podiatr Med 1998;32:55–9.
18. Bennett PJ, Patterson C, Wearing S, Baglioni T. Development and validation of a questionnaire designed to measure foot-health status. J Am Podiatr Med Assoc 1998;88:419–28.
19. Williams AE, Rome K, Nester CJ. A clinical trial of specialist footwear for patients with rheumatoid arthritis. Rheumatology (Oxford) 2007;46: 302–7.
20. Burns J, Crosbie J, Ouvrier R, Hunt A. Effective orthotic therapy for the painful cavus foot: a randomized controlled trial. J Am Podiatr Med Assoc 2006;96:205–11.
21. Rome K, Gray J, Stewart F, Hannant SC, Callaghan D, Hubble J. Evaluating the clinical effectiveness and cost-effectiveness of foot orthoses in the treatment of plantar heel pain: a feasibility study. J Am Podiatr Med Assoc 2004;94:229–38.
22. Dunn JE, Link CL, Felson DT, Crincoli MG, Keysor JJ, McKinlay JB. Prevalence of foot and ankle conditions in a multiethnic community sample of older adults. Am J Epidemiol 2004;159:491–8.
23. Beeson P, Phillips C, Corr S, Ribbans WJ. Hallux rigidus: a cross-sectional study to evaluate clinical parameters. Foot (Edinb) 2009;19: 80–92.
24. Maher AJ, Metcalfe SA. First MTP joint arthrodesis for the treatment of hallux rigidus: results of 29 consecutive cases using the foot health status questionnaire validated measurement tool. Foot (Edinb) 2008; 18:123–30.
25. Radford JA, Landorf KB, Buchbinder R, Cook C. Effectiveness of low-Dye taping for the short-term treatment of plantar heel pain: a randomised trial. BMC Musculoskelet Disord 2006;7:64.
26. Ferreira AF, Laurindo IM, Rodrigues PT, Ferraz MB, Kowalski SC, Tanaka C. Brazilian version of the foot health status questionnaire (FHSQ-BR): cross-cultural adaptation and evaluation of measurement properties. Clinics (Sao Paulo) 2008;63:595–600.
27. Sirera-Vercher MJ, Saez-Zamora P, Sanz-Amaro MD. Translation, transcultural adaptation to Spanish, to Valencian language of the Foot Health Status Questionnaire. Revista Espanola de Cirugia Ortopedica y Traumatologia (English Edition) 2010;54:211–9.
28. British Psychological Society Steering Committee on Test Standards. Psychological testing: a user's guide. Leicester (UK): The British Psychological Association; 1995.
29. Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, editors. Testing structural equation models. Beverly Hills (CA): Sage Publishing; 1993.
30. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999;6:1–55.
31. Badlissi F, Dunn JE, Link CL, Keysor JJ, McKinlay JB, Felson DT. Foot musculoskeletal disorders, pain, and foot-related functional limitation in older persons. J Am Geriatr Soc 2005;53:1029–33.
32. Tully MP, Cantrill JA. Subjective outcome measurement: a primer. Pharm World Sci 1999;21:101–9.
33. World Health Organization. Towards a common language for functioning, disability and health: ICF. URL: http://www.who.int/classifications/icf/site/beginners/bg.pdf.
34. Garrow AP, Papageorgiou AC, Silman AJ, Thomas E, Jayson MI, Macfarlane GJ. Development and validation of a questionnaire to assess disabling foot pain. Pain 2000;85:107–13.
35. Menz HB, Tiedemann A, Kwan MM, Plumb K, Lord SR. Foot pain in community-dwelling older people: an evaluation of the Manchester Foot Pain and Disability Index. Rheumatology (Oxford) 2006;45:863–7.
36. Dawson J, Coffey J, Doll H, Lavis G, Cooke P, Herron M, et al. A patient-based questionnaire to assess outcomes of foot surgery: validation in the context of surgery for hallux valgus. Qual Life Res 2006;15: 1211–22.
37. Cook CE, Cleland J, Pietrobon R, Garrow AP, Macfarlane GJ. Calibration of an item pool for assessing the disability associated with foot pain: an application of item response theory to the Manchester Foot Pain and Disability Index. Physiotherapy 2007;93:89–95.
38. Garrow AP, Silman AJ, Macfarlane GJ. The Cheshire Foot Pain and Disability Survey: a population survey assessing prevalence and associations. Pain 2004;110:378–84.
39. Waxman R, Woodburn H, Powell M, Woodburn J, Blackburn S, Helliwell P. FOOTSTEP: a randomized controlled trial investigating the clinical and cost effectiveness of a patient self-management program for basic foot care in the elderly. J Clin Epidemiol 2003;56:1092–9.
40. Kaoulla P, Frescos N, Menz HB. Development and validation of a Greek language version of the Manchester Foot Pain and Disability Index. Health Qual Life Outcomes 2008;6:39.
41. Marinozzi A, Martinelli N, Panasci M, Cancilleri F, Franceschetti E, Vincenzi B, et al. Italian translation of the Manchester-Oxford Foot Questionnaire, with re-assessment of reliability and validity. Qual Life Res 2009;18:923–7.
42. Ferrari SC, Cristina dos Santos F, Guarnieri AP, Salvador N, Abou Hala Correa AZ, Abou Hala AZ, et al. Indice Manchester de incapacidade associada ao pe doloroso no idoso: traducao, adaptacao cultural e validacao para a lingua portuguesa. Revista Brasileira de Reumatologia 2008;48:335–41.
43. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
44. King G, Murray CJ, Salomon JA, Tandon A. Enhancing the validity and cross-cultural comparability of measurement in survey research. Am Polit Sci Rev 2004;98:191–207.
45. Williams A, Nester C, Ravey M. Rheumatoid arthritis patients' experiences of wearing therapeutic footwear: a qualitative investigation. BMC Musculoskelet Disord 2007;8:104.
46. Roddy E, Muller S, Thomas E. Defining disabling foot pain in older adults: further examination of the Manchester Foot Pain and Disability Index. Rheumatology (Oxford) 2009;48:992–6.
47. Macran S, Kind P, Collingwood J, Hull R, McDonald I, Parkinson L. Evaluating podiatry services: testing a treatment specific measure of health status. Qual Life Res 2003;12:177–88.
48. Farndon L, Barnes A, Littlewood K, Harle J, Beecroft C, Burnside J. Clinical audit of core podiatry treatment in the NHS. J Foot Ankle Res 2009;2:7.
49. Rowan K. The development and validation of a multi-dimensional measure of chronic foot pain: the Rowan Foot Pain Assessment Questionnaire (ROFPAQ). Foot Ankle Int 2001;22:795–809.
50. Melzack R, Casey KL. Sensory, motivational and central control determinants of chronic pain: a new conceptual model. The Skin Senses 1968:423–43.

Summary Table for Self-Administered Patient/Participant-Reported Foot Health Questionnaires*

Scale	Purpose/content	Number of items	Subscales (no. questions)	Reliability evidence	Validity evidence	Ability to detect change	MDD	MCID
AAOS-FAM	Evaluate patient perception of foot health and measure surgical outcomes	25 questions in 5 subscales	Pain (9 questions); foot function (6 questions); stiffness and swelling (2 questions); giving way (3 questions); shoe comfort (5 questions)	Good	Good	Good	Unknown	Unknown
BFS	Assess patient perception of impact of foot problems on everyday life	15 questions in 3 subscales	Foot concern and pain (7 questions); footwear and general foot health (4 questions); mobility (3 questions)	Adequate	Good	Good	Unknown	Unknown
FFI-R	Assess foot-related health and quality of life	34 or 68 questions, with 68 questions (long form) having 4 subscales (items in this table are for long form)	Pain and stiffness (19 questions); social and emotional outcomes (19 questions); disability (20 questions); activity limitation (10 questions)	Excellent	Good	Unknown	Unknown	Unknown
FHSQ	Measure foot health related to quality of life	13 questions in 4 subscales	Foot pain (4 questions); foot function (4 questions); footwear (3 questions); general foot health (2 questions)	Good	Excellent	Excellent	Excellent	Excellent
MFPDI	Measure disabling foot pain in general population	17 questions in 3 subscales	Functional limitation (10 questions); pain intensity (5 questions); perception of foot appearance (2 questions)	Good	Good	Unknown	Unknown	Unknown
PHQ	Measure foot-related health in podiatric patient populations	7 items (1 question in 6 subscales and visual analog scale)	1 question each for: walking, foot hygiene, nail care, foot pain, worry about feet, quality of life. Visual analog scale for current foot status	Unknown	Good	Good	Unknown	Unknown
ROFPAQ	Evaluate chronic foot pain	39 questions in 3 subscales and comprehension questions	Sensory (16 questions); affective (10 questions); cognitive (10 questions)	Excellent	Good	Unknown	Unknown	Unknown

* MDD = minimum detectable difference; MCID = minimum clinically important difference; AAOS-FAM = American Academy of Orthopedic Surgeons Lower Limb Outcomes Assessment: Foot and Ankle Module; BFS = Bristol Foot Score; FFI-R = Revised Foot Function Index; FHSQ = Foot Health Status Questionnaire; MFPDI = Manchester Foot Pain and Disability Index; PHQ = Podiatric Health Questionnaire; ROFPAQ = Rowan Foot Pain Assessment.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S37–S46
DOI 10.1002/acr.20572
© 2011, American College of Rheumatology

Measures of Adult Systemic Lupus Erythematosus

Updated Version of British Isles Lupus Assessment Group (BILAG 2004), European Consensus Lupus Activity Measurements (ECLAM), Systemic Lupus Activity Measure, Revised (SLAM-R), Systemic Lupus Activity Questionnaire for Population Studies (SLAQ), Systemic Lupus Erythematosus Disease Activity Index 2000 (SLEDAI-2K), and Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index (SDI)

JUANITA ROMERO-DIAZ,[1] DAVID ISENBERG,[2] AND ROSALIND RAMSEY-GOLDMAN[1]

INTRODUCTION

Measurement of disease activity in systemic lupus erythematosus (SLE) is central to evaluating outcomes, differences among SLE patient groups, responses to a new drug proposed, and also for assessing disease longitudinally for observational and clinical trials. Several validated and updated instruments have been available since the early 1980s, but more recent studies gauging reliability and validity for classifying and monitoring groups of patients in the research setting are now available.

Two cardinal features of SLE have challenged investigators refining these tools: first, the complex multisystem nature of this disease with fluctuating levels of disease activity, which may vary between patients and within the same patient over time; second, the absence of a "gold standard" for determining the psychometric properties of each proposed scale limits comparisons to expert opinion using a physician's visual analog scale or by comparing one scale against other to assess performance across proposed instruments. However, these strategies do not eliminate bias based on personal experience, nor do they differentiate between different opinions on the relative importance of disease manifestations in different systems.

Therefore, an experience-based evaluation may be subject to greater interrater variability than the use of the disease activity instrument itself. Furthermore, psychometric properties should be influenced by the length of the scale (number of items and scoring scale), number of patients included, or disease severity of patients under study.

Two main types of activity measures in SLE have been developed: global score systems (for example, the European Consensus Lupus Activity Measurements, Systemic Lupus Activity Measure [SLAM], and Systemic Lupus Erythematosus Disease Activity Index [SLEDAI]), which provide an overall measure of activity, and individual organ/system assessment scales that assess disease activity in single organs (such as the British Isles Lupus Assessment Group Index [BILAG]). The Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index score is a measure for chronic damage; it has been included due to its prognostic value in clinical and research basis.

The SLEDAI, SLAM, and BILAG have performed in effective and reliable manners in studies; furthermore, they correlate with one another (1–3). The SLEDAI, Safety of Estrogens in Lupus Erythematosus National Assessment (SELENA)–SLEDAI, SLEDAI 2000 (4–7), and BILAG (8–10) have been successfully used in observational trials and case studies, although baseline disease activity index (DAI) scores were not always predictors of subsequent damage or other outcomes (11,12). These DAIs were validated in the context of long-term observational trials studies and not in randomized clinical trials (RCTs) (1,9,10,13–15). The few RCTs conducted have shown that improvement in DAI scores correlates with response rates, disease remission, and flare prevention; however, a threshold of clinically meaningful change has not been established (1,13,16,17). Current work has focused on developing a responder index developed in collaboration with the Food and Drug Administration–defined response as improvement and/or no deterioration in patient- and

Supported in part by the Institute of Science and Technology at Mexico City (grant BM09-209) and the NIH (grants K24-AR-002138 and P60-2-AR-30692).

[1]Juanita Romero-Diaz, MD, MS, Rosalind Ramsey-Goldman, MD, DrPH: Northwestern University Feinberg School of Medicine, Chicago, Illinois; [2]David Isenberg, MD: University College London, London, UK.

Address correspondence to Rosalind Ramsey-Goldman, MD, DrPH, Division of Rheumatology, Northwestern University Feinberg School of Medicine, 240 East Huron, Suite M300, Chicago, IL 60611. E-mail: rgramsey@northwestern.edu.

Submitted for publication February 1, 2011; accepted in revised form July 28, 2011.

physician-reported outcomes. The SLE responder index, which utilizes the SELENA–SLEDAI score to determine global improvement, BILAG domain scores to ensure no significant worsening in heretofore unaffected organ systems, and physician's global assessment to ensure that improvements in disease activity are not achieved at the expense of the patient's overall condition, which may have been missed by either DAI, is one example used in a recent clinical trial (18). Ongoing work to refine or develop responder indices will enhance our ability to measure meaningful outcomes in future RCTs.

For purpose of this review, we selected those indices that have shown the strongest evidence of validity when used by investigators from different countries in large studies of patients with SLE. The exact choice of instrument should be governed by the purpose for which it is required in clinical practice or research.

UPDATED VERSION OF BRITISH ISLES LUPUS ASSESSMENT GROUP (BILAG 2004)

Description

Purpose. To assess lupus activity based upon the "intent-to-treat" premise. The original version was published in 1988 (19). Over time, several deficiencies were noted by members of BILAG, which prompted a major revision. The updated version (BILAG 2004) was published in 2005 (20).

Content. Specific manifestation in 9 systems. In this revised index, the original section of vasculitis has been removed and 2 systems were added: ophthalmic and abdominal.

Number of items. 101 and 5 additional items required mainly for calculations of glomerular filtration rate.

Response options/scale. Each question is answered as: 0 = not present, 1 = improving, 2 = same, 3 = worse, and 4 = new.

Recall period for items. It records disease activity occurring over the past 4 weeks as compared with the previous 4 weeks.

Endorsements. Adult patients with systemic lupus erythematosus (SLE).

Examples of use. Yee CS, Isenberg DA, Prabu A, Sokoll K, Rahman A, Bruce IN, et al. BILAG 2004 index captures systemic lupus erythematosus disease activity better than SLEDAI-2000. Ann Rheum Dis 2008;67:873–6 (21).

Isenberg DA, Allen E, Farewell V, D'Cruz D, Alarcon GS, Aranow C, et al. An assessment of disease flare in patients with systemic lupus erythematosus: a comparison of BILAG 2004 and the flare version of SLEDAI. Ann Rheum Dis 2011;70:54–9 (22).

Practical Application

How to obtain. The BILAG 2004 index assessment and BILAG 2004 index glossary can be obtained at *Rheumatology* online as supplementary data without cost.

Contact information. The BILAG group: current chair of the BILAG group is Professor David Isenberg, Room 331, The Windeyer Building, University College London, 46 Cleveland Street, London W1T 4JF, UK.

Method of administration. Physician completed.

Training. Formal training of raters and a well-defined glossary are needed.

Equipment needed. None to complete the index. To calculate categorical or numerical scoring, a computer program is needed.

Scoring. As above, each question is recorded as 0, 1, 2, 3, or 4. Then, a computer program facilitates scoring from numerical to alphabetical score for each system (grade A–E).

Score interpretation. The BILAG 2004 index categorizes disease activity into 5 different levels from A–E. Grade A represents very active disease likely necessitating immunosuppressive drugs and/or a prednisolone (or equivalent) dose of >20 mg daily or high-dose anticoagulation. Grade B represents moderate disease activity requiring a lower dose of corticosteroids, topical steroids, topical immunosuppressive drugs, antimalarials, or nonsteroidal antiinflammatory drugs. Grade C indicates mild stable disease, while grade D implies no disease activity but the system had previously been affected. Grade E indicates no current or previous disease activity.

Respondent burden. 5–20 minutes, plus time to complete history and physician examination.

Administrative burden. Up to 50 minutes. The instrument cannot be scored until laboratory results are available, and this may take a few days.

Translations/adaptations. English only. The original BILAG index is available in computer version.

Psychometric Information

Method of development. The BILAG 2004 was developed by nominal consensus. The members of the BILAG developed the BILAG index by making agreed assumptions about the likely treatment that will be given to patients with particular groups of clinical features. There was no attempt to weight the importance of involvement of different systems. Items were generated by detailed discussion of BILAG members.

Reliability. Good reliability (intraclass correlation coefficient [ICC] >0.60) and high levels of physician agreement ($\sigma_{\text{physician}}/\sigma_{\text{patient}} = <0.40$) were shown in 2 real-patient exercises.

The interrater reliability of the index was shown in a multicenter study of 97 "live" patients in 2 exercises (E1 and E2). The overall ICC determined in E1 was 0.45 (95% confidence interval [95% CI] 0.31, 0.58), and in E2 was 0.67 (95% CI 0.54, 0.76). There was improvement in the levels of agreements and in the kappa and ICC reliability from E1 to E2. Four items with poor agreement between raters were identified. Training of raters was suggested to ensure the optimal performance of the index (23).

Validity. In a multicenter cross-sectional study of 369 patients, scores indicating active disease (overall BILAG 2004 scores of A and B) were significantly associated with increase in therapy (odds ratio 19.3, $P < 0.01$). The overall sensitivity of the index was 81%, specificity was 81.9%, positive predictive value was 56.8%, and negative predictive value was 93.6%. Construct and criterion validity were also shown (24).

Ability to detect change. Using the external method responsiveness, the BILAG 2004 has been shown to be sensitive to change to assess SLE disease activity. This has been shown in a longitudinal study that involved 8 centers in the UK in which 1,761 visits from 347 SLE patients were evaluated. Increase in the overall score was associated with increase in therapy (coefficient 1.35; 95% CI 1.01, 1.70) and inversely associated with decrease in therapy (coefficient −0.44; 95% CI −0.81, −0.06) (25).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. This score incorporates the important element of change in disease state with time (the delta factor). It is sensitive to small changes and distinguishes between disease activity and disease severity. It is the only validated lupus activity index that aims to show activity in individual systems "at a glance" rather than combining them into a global score.

Caveats and cautions. Formal training of raters and a well-defined glossary are essential to ensure the optimal performance of the index.

Clinical usability. The BILAG 2004 index was developed particularly for research. However, it should be useful to monitor the disease for individuals due its ability to identify whether the disease is improving, stable, or worsening.

Research usability. The BILAG 2004 index is appropriate for investigations of disease outcome and treatment protocols. Despite the complex calculations, the score is quick to conduct, especially when calculated by a computer, and only minimally dependent on the particular clinician carrying out the procedure. To facilitate comparisons with global indices, a numerical scoring system has been associated with the BILAG 2004 index. The optimal method is to convert the assessments so that an "A" = 12 points, "B" = 8 points, "C" = 1 point, and "D/E" = 0 points (26).

EUROPEAN CONSENSUS LUPUS ACTIVITY MEASUREMENTS (ECLAM)

Description

Purpose. To assess disease activity in patients with lupus within the past month.

Content. Lupus activity is divided into 10 organs/systems, plus erythrocyte sedimentation rate (ESR) and complement levels with varying numbers of items in each. Emphasis is on evolving changes.

Number of items. 33 items.

Response options/scale. There are 12 categories (10 organs/systems plus ESR and complement levels), 4 of which are divided into subcategories.

Recall period for items. The last month.

Endorsements. Disease activity in patients with systemic lupus erythematosus (SLE).

Examples of use. American College of Rheumatology Ad Hoc Committee on Systemic Lupus Erythematosus Response Criteria. The American College of Rheumatology response criteria for systemic lupus erythematosus clinical trials: measures of overall disease activity. Arthritis Rheum 2004;50:3418–26 (27).

Mosca M, Chimenti D, Pratesi F, Baldini C, Anzilotti C, Bombardieri S, et al. Prevalence and clinico-serological correlations of anti-α-Enolasa, anti-C1q, and anti-dsDNA antibodies in patients with systemic lupus erythematosus. J Rheumatol 2006;33:695–7 (28).

Amital H, Szekanecz Z, Szucz G, Danko K, Nagy E, Csepany T, et al. Serum concentration of 25-OH vitamin D in patients with systemic lupus erythematosus (SLE) are inversely related to disease activity: is it time to routinely supplement patients with SLE with vitamin D? Ann Rheum Dis 2010;69:1155–7 (29).

Practical Application

How to obtain. *Contact information.* The European Workshop for Rheumatology Research. Main developer and contact person is Professor Stephano Bombardieri, Universidad of Pisa, Italy.

Method of administration. Physician completed.

Scoring. Simple additive.

Score interpretation. Range is 0–17.5. This is a global score index. Item scores range from 0.5 (e.g., fever/fatigue) to 2 (e.g., new neuropsychiatric/evolving renal manifestation).

Respondent burden. Up to 10 minutes.

Administrative burden. A history and physician examination is needed. For a reasonably stable patient, <5 minutes; for a complicated patient, up to 10 minutes. Training is needed, especially in a multicenter studies.

Translations/adaptations. English and Italian versions available. Paper or computer versions.

Psychometric Information

Method of development. The ECLAM was constructed during the course of a multicenter study involving 704 patients, on the basis of the correlations found for each patient between a wide range of clinical/laboratory parameters with the clinician's assessment of disease activity (the gold standard). Multivariate regression analyses were carried out to evaluate the combined performance of different sets of clinical and serologic variables in predicting disease activity, and to define the relative weight of each variable in terms of regression coefficients in multivariate models (30).

Reliability. Data from 32 consecutive patients were obtained from 4 observers (2 experts, 1 trainee, 1 nurse). The correlation coefficients between ECLAM scores ranged from 0.9–0.95 (31). In a second study, 64 consecutive patients were scored at time of evaluation and 2 weeks later from chart data by 2 observers. The correlation coefficient between patient and chart ECLAM score was 0.88 and the interobserver variability was low, with a correlation coefficient ranging from 0.9–0.93 (32).

Validity. Data from 75 patients (19 centers) were collected and each patient was observed twice over 3 months. The ECLAM index at each time point was compared with the Systemic Lupus Activity Measure (SLAM), Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), and British Isles Lupus Assessment Group (BILAG). The correlation coefficients for the ECLAM compared with the others indices ranged from 0.72–0.78 (3).

Ability to detect change. In 23 patients seen every 2 weeks for up to 40 weeks, 5 disease activity measures were completed along with the physician's and patient's global assessments. Changes in SLE activity were correlated with each activity measure, and for the ECLAM, r = 0.65. Sensitivity to change was greatest for the ECLAM when compared with the physician's global assessment. Using a standardized response measure, the score for the ECLAM was 0.75 (3).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The ECLAM index was directly derived from a large number of real patients and the analysis of a large amount data collated in a standardized manner during a multicenter study.

Caveats and cautions. Global score will miss changes in severity over time.

Clinical usability. The ECLAM index should be an excellent tool for clinical usability because of its great simplicity. It is based on 12 of the most common parameters of disease activity.

Research usability. The ECLAM score has been widely used in sets of real and paper patient exercises mostly comparing it with the SLEDAI, SLAM, and BILAG. It has been shown to be a reliable instrument for calculating disease activity retrospectively from clinical charts when used in the setting of a tertiary center for patient care.

SYSTEMIC LUPUS ACTIVITY MEASURE, REVISED (SLAM-R)

Description

Purpose. To measure the degree of disease activity in patients with systemic lupus erythematosus (SLE) within the last month. It was published in 1988 and revised in 1991 (33).

Content. Specific manifestation in 9 organs/systems, plus 7 laboratory features.

Number of items. 9 organs/systems, with laboratory category.

Response options/scale. Organ items scored 0–3 points if present within the last month (severity incorporated into higher score per item). Most items can score a maximum of 3 points. Few items can score a maximum of 1 point. The laboratory category can score a maximum of 21 points.

Recall period for items. The SLAM covers symptoms that occurred during the previous month.

Endorsements. Patients with SLE.

Examples of use. Chang ER, Abrahamowics M, Ferland D, Fortin PR. Organ manifestations influence differently the responsiveness of 2 lupus disease activity measures, according to patients' or physicians' evaluations of recent lupus activity. J Rheumatol 2002;29:2350–8 (34).

Zhang J, Gonzales LA, Roseman JM, Vila LM, Reveille JD, Alarcon GS. Predictors of the rate of change in disease activity over time in LUMINA, a multiethnic US cohort of patients with systemic lupus erythematosus: LUMINA LXX. Lupus 2010;19:727–33 (35).

Practical Application

How to obtain. Copyrighted by Fellows of Harvard College; developer and contact person is Dr. Matthew Liang, Professor of Medicine, Department of Medicine/Rheumatology/Immunology, PBB-82, Brigham & Women's Hospital, 75 Francis Street, Boston, MA 02115. The computer version is available from Gordon Hamilton (e-mail: LIMATHON@aol.com).

No cost to use (unless the computerized version is needed, then cost depends upon type of usage [commercial/academic]).

Method of administration. Physician completed. Questionnaire available in paper format (optical scannable) or as part of the BLIPS software program.

Scoring. Simple additive.

Score interpretation. *Score range.* Maximum score is 81 points. Judgment as to whether manifestations (laboratory or otherwise) are due to lupus is needed. A score of 7 is considered clinically important and effects decision to treat.

Respondent burden. Up to 15 minutes.

Administrative burden. A complete history and physical examination is needed. To complete the form in an essentially well patient with a short history takes <10 minutes. For a complex patient not well known to the physician it can take up to 15 minutes. For most patients it takes <10 minutes.

Training is needed to develop consensus on subjective components of the index, especially in multicenter studies. Dr. Matthew Liang (contact information above) or Dr. Paul Fortin (Division of Rheumatology, Room MP-10-304, Toronto Western Hospital, 399 Bathurst Street, Toronto, Ontario M5T 2S8, Canada) is suggested.

Translations/adaptations. Available in English, Korean, German, and Chinese.

Psychometric Information

Method of development. It was developed based on domain sampling theory. Items chosen for the scale represent those manifestations that occur more frequently, those that can be graded, and those that can be operationally defined and reliably rated.

Reliability. The reliability of the index was shown in a study of 25 "live" patients seen twice over a 3–5-week period and 2 physicians who were not providing care for the patients. The SLAM index interrater reliability and intervisit reliability were 0.86 and 0.73, respectively.

The reliability of the SLAM-R was demonstrated in a study of 30 patients seen twice 2–4 weeks apart by 2 physicians who were not providing care for the patient.

The SLAM-R index interrater reliability and intervisit reliability were 0.78 and 0.85, respectively (36).

Validity. *Convergent and discriminant.* The validity of the index was shown in a study of 25 "live" patients seen twice over a 3–5-week period and 2 physician raters using 6 scales, including the SLAM. These raters were not providing care for the patients. The average correlation between the SLAM and the other scales was 0.9, ranging from 0.9–1.0. Furthermore, when correlations were evaluated to assess change between visits, the range was 0.5–0.8 across instruments, demonstrating convergent validity. The various components contributing to the total variance of the SLAM were 73% for patients, 13% for visits, and 14% for raters demonstrating discriminant validity (37).

Construct validity of the SLAM-R. The correlation between the SLAM-R scores, the physician's global assessment, anti–double-stranded DNA, C3, and C4 were statistically significant, ranging from −0.29 to 0.87 (37).

Ability to detect change. Excellent sensitivity and responsiveness to change have been shown in comparative studies with the British Isles Lupus Assessment Group (BILAG) and Systemic Lupus Erythematosus Disease Activity Index (SLEDAI). In an international validation study, where 8 patients with 3 visits were rated by the Systemic Lupus International Collaborating Clinics group using 3 indices (SLEDAI, BILAG, SLAM), all indices were able to detect differences between patients ($P < 0.01$) (38).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. This index includes both dimensions: disease activity and disease severity.

Caveats and cautions. One of its disadvantages is that many items are subjective, because scoring relies on the reporting of symptoms by the patients rather than objective documentation. Difficulty in distinguishing changes, i.e., patients with multiple mild or improving manifestations compared to those with 1 or 2 severe features. Note that some of the most severe items also count as damage, i.e., cerebrovascular accident.

Clinical usability. For this index, a score of ≥7 is considered clinically important because it is associated with a probability of initiating therapy in >50% of cases. However, it is important to consider that it gives equal weighting to mild and serious organ disease activity without considering the significance of the organ involved.

Research usability. This index has a high sensitivity to change and responsiveness when the patient's global assessment is considered as the standard. The SLAM correlates with several aspects of the patient's perception of health, as evaluated with the Short Form 36 (34,35,39).

SYSTEMIC LUPUS ACTIVITY QUESTIONNAIRE FOR POPULATION STUDIES (SLAQ)

Description

Purpose. To provide an economic way of following and tracking disease activity for large groups of systemic lupus erythematosus (SLE) patients who may be at a distance from a center in epidemiologic studies. It was developed based on items from the Systemic Lupus Activity Measure (SLAM) (40). It was published in 2003.

Content. Specific symptoms of disease activity and a single numerical rating scale (NRS) asking the patient to rate disease activity on a scale of 0–10 over the past 3 months.

Number of items. 24 items in 9 organs/systems weighted and aggregated in a manner analogous to the scoring system used in the SLAM.

Response options/scale. For questions regarding disease activity, there are 4 options, as follows: no problem = 0, mild = 1, moderate = 2, and severe = 3. For a single NRS, it rates from 0 = "no activity" to 10 = "most activity."

Recall period for items. The last 3 months.

Endorsements. Studies with large groups of SLE patients.

Examples of use. Trupin L, Tonner MC, Yazdany J, Julian LJ, Criswell LA, Katz PP, et al. The role of neighborhood and individual socioeconomic status in outcomes of systemic lupus erythematosus. J Rheumatol 2008;35: 1782–8 (41).

Wolfe F, Petri M, Alarcon GS, Goldman J, Chakravaty EF, Katz RS, et al. Fibromyalgia, systemic lupus erythematosus (SLE), and evaluation of ALE activity. J Rheumatol 2009;36:82–8 (42).

Practical Application

How to obtain. A copy can be obtained for 1 study (see Appendix A) (40) without cost.

Method of administration. Patient self-completed questionnaire or telephone administration.

Scoring. Arithmetic computation by hand.

Score interpretation. Scores can range from 0–44. It correlates with the physician-completed SLAM.

Respondent burden. Up to 10 minutes.

Administrative burden. Up to 10 minutes.

Translations/adaptations. English only. No adaptations available.

Psychometric Information

Method of development. It was developed based on domain sampling theory. Under a clinical setting, assessments of 93 patients who presented to an academic medical center for clinical care were used. It was based on items from the SLAM that are amenable to self-report (40).

Reliability. In an observational cohort study of 982 English-speaking patients with SLE, the SLAQ demonstrated excellent internal consistency, with a Cronbach's α of 0.87. Data structure examined by principal factor analysis showed that 1 factor accounted for 92% of the variance (43).

Validity. Construct validity was demonstrated by examining correlation of the SLAQ with measures that are likely to be related to disease activity in SLE (r = 0.51–0.73) (43).

Ability to detect change. The SLAQ demonstrated a small to moderate degree of responsiveness for participants who reported a perceived change in disease status;

standardized response means were 0.66 and −0.37 for those reporting clinical worsening and improvement, respectively (43).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SLAQ index is a unique instrument developed and validated for measure disease status outside the clinical setting in SLE patients. It is very useful for large epidemiologic studies in which many patients live outside the catchment area or physician-directed assessment may prove impractical and costly.

Caveats and cautions. The SLAQ instrument should not be used instead of careful clinical followup of patients in day-to-day practice. If the level of education may influence, the response rate needs to be evaluated. Future studies are needed to confirm the reliability of the SLAQ compared with a physician assessment, particularly in different age, sex, and racial/ethnic groups.

Clinical usability. The SLAQ is intended to be used as an initial screen to identify subjects with new or increased disease activity who need further evaluation by a physician (positive predictive value ranged from 56–89% for detecting clinically significant disease activity).

Research usability. The SLAQ demonstrated adequate reliability, construct validity, and responsiveness in a large community-based cohort of patients with SLE.

SYSTEMIC LUPUS ERYTHEMATOSUS DISEASE ACTIVITY INDEX 2000 (SLEDAI-2K)

Description

Purpose. To measure disease activity in patients with systemic lupus erythematosus (SLE). The original version was introduced in 1985 (15,44). In 2002, it was modified to reflect persistent active disease in those descriptors that had previously considered new or recurrent occurrences (SLEDAI-2K) (45).

Content. Specific manifestation in 9 organs/systems.

Number of items. 24 items covering 9 organs systems.

Response options/scale. There are 24 items for the 9 organs/systems. Scored if present within the last 10 days. Two systems can score a maximum of 8 points each, 2 systems can score a maximum of 4 points each, 3 systems can score a maximum of 2 points each, and 2 systems can score a maximum of 1 point each. Scores range from 0–105 points.

Recall period for items. Disease activity within the last 10 days. Recently, the SLEDAI-2K for a timeframe of 30 days prior to a visit for clinical and laboratory variables was shown to be similar to the SLEDAI-2K for 10 days (46).

Endorsements. Disease activity in patients with SLE.

Examples of use. Uribe AG, Vila LM, McGwin G Jr, Sanchez ML, Reveille JD, Alarcon GS. The Systemic Lupus Activity Measure-Revised, the Mexican Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), and a modified SLEDAI-2K are adequate instruments to measure disease activity in systemic lupus erythematosus. J Rheumatol 2004;31:1934–40 (47).

Petri M, Kim MY, Kalunian KC, Grossman J, Hahn BH, Sammaritano LR, et al. Combined oral contraceptives in women with systemic lupus erythematosus. N Engl J Med 2005;353:2550–8 (48).

Sanchez-Guerrero J, Uribe AG, Jimenez-Santana L, Mestanza-Peralta M, Lara-Reves P, Seuc AH, et al. A trial of contraceptive methods in women with systemic lupus erythematosus. N Engl J Med 2005;353:2539–49 (49).

Practical Application

How to obtain. The Toronto Group: Claire Bombardier, MD, initial development only), Drs. Dafna Gladman, MD, and Murray Urowitz, MD (Toronto Western Hospital, 399 Bathurst Street IE – 410B, Toronto, Ontario, Canada M5T 2S8).

Method of administration. Physician completed.

Scoring. Simple additive.

Score interpretation. The score range is 0–105 points. A score of 6 is considered clinically important and affects decision to treat.

Respondent burden. Up to 10 minutes.

Administrative burden. A complete history and physical examination is needed. The instrument cannot be scored until laboratory results are available, and this may take a few days. To complete the form in an essentially well patient with a short history it can take <10 minutes. For a complex patient not well known to the physician it can take <10 minutes.

Translations/adaptations. The SLEDAI-2K is available in English and Spanish. Some adaptations have been published, e.g., the Safety of Estrogens in Lupus Erythematosus National Assessment–SLEDAI used in the Safety of Estrogen trial. It was modified from the SLEDAI to insure that the descriptors of organ system involvement reflected ongoing disease activity (50,51). The Mexican modification of the SLEDAI, a simplified version without the immunologic test, makes the index cheaper to administer (52).

Psychometric Information

Method of development. It was developed with a panel of experienced rheumatologists with expertise in SLE, using well-established group techniques and index development methodology.

Reliability. The reliability of the original SLEDAI was shown in a paper patient exercise in which 534 scenarios were generated from real patient data and 14 lupus experts participated in an interrater reliability study. The interrater correlation for the 46 most common patients profiles ranged from 0.61–0.80 (15).

The reliability of the SLEDAI-2K was evaluated in a multicenter multiethnic study where 93 patients were studied. Agreement for each of the items was between 81.7% and 100% (10).

Validity. A group of 14 lupus experts completed a testing set of 69 real scenarios with common manifestations, 98 anchor profiles, and 116 real patient cases. The intraclass correlation coefficient was 0.79, representing slightly stronger agreement within cases with common manifestations of disease than for unique (0.71) or anchor profiles (0.64) (15).

The SLEDAI-2K was validated against the SLEDAI using all visits in a cohort of 960 patients in the Toronto databank; there was a high correlation between both indices (r = 0.97, $P = 0.0001$) (45).

Ability to detect change. The SLEDAI sensitivity and responsiveness to change have been shown in comparative studies with the Systemic Lupus Activity Measure, British Isles Lupus Assessment Group, and European Consensus Lupus Activity Measurements. In a prospective study, 23 patients with SLE were examined every 2 weeks for up to 40 weeks. Estimates of sensitivity to change varied with the standard used. The sensitivity to change was smallest for the SLEDAI, with a standardized response mean (SRM) of 0.48 when the physician global assessment was used as the standard and an SRM of −0.01 when the patient global assessment was used (3,38).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. All versions are validated and used by lupus researchers for clinical and research purposes.

Caveats and cautions. The SLEDAI does not record improving or worsening, and does not include severity within an organ system.

Clinical usability. Activity categories have been defined on the basis of the SLEDAI score. A SLEDAI score >5 is associated with a probability of initiating therapy in >50% of cases.

Research usability. Neither version of the SLEDAI captures improving or worsening. This probably explains why it is less sensitive to change than other instruments.

SYSTEMIC LUPUS INTERNATIONAL COLLABORATING CLINICS (SLICC)/ AMERICAN COLLEGE OF RHEUMATOLOGY DAMAGE INDEX (SDI)

Description

Purpose. To capture those items of permanent change that has occurred in patients after a diagnosis of systemic lupus erythematosus (SLE), regardless of attribution.

Content. Specific manifestation in 12 organ systems.

Number of items. 41 items covering 12 organ systems. Within each scale or system, a variable number of components are to be found (up to 6).

Response options/scale. Thirty-one items score 1 point if present. Six items can score a maximum of 2 points; 1 item can score a maximum of 3 points.

Recall period for items. Duration of manifestation (or irreversibility), i.e., must be present for a minimum of 6 months or expected not to reverse, such as surgical procedure or infarction.

Endorsements. Measure damage in patients with SLE.

Examples of use. Gladman D, Ginzler E, Goldsmith C, Fortin P, Liang M, Urowitz M, et al. The development and initial validation of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index for systemic lupus erythematosus. Arthritis Rheum 1996;39:363–9 (53).

Stoll T, Seifert B, Isenberg DA. SLICC/ACR Damage Index is valid and renal and pulmonary organ scores are predictors of severe outcome in patients with systemic lupus erythematosus. Br J Rheumatol 1996;35:248–54 (54).

Rahman P, Gladman DD, Urowitz MB, Hallett D, Tam LS. Early damage as measured by the SLICC/ACR Damage Index is a predictor of mortality in systemic lupus erythematosus. Lupus 2001;10:93–6 (11).

Practical Application

How to obtain. *Contact information.* Dr. Dafna Gladman, Toronto Western Research Institute, University Health Network, Toronto Western Hospital, 399 Bathurst Street, IE-410B Toronto, Ontario, Canada M5T 2S8.

Questionnaire available in paper format or as part of the BLIPS software program. The computer version is available from Gordon Hamilton (e-mail: LIMATHON@aol.com).

Method of administration. Physician completed.

Scoring. As above, the duration of manifestation (or irreversibility), i.e., must be present for a minimum of 6 months or expected not to reverse, such as surgical procedure or infarction. Item scored regardless of attribution to SLE; therefore, this catches morbidity from treatment from SLE or other complications that may be increased in SLE, e.g., fracture, etc.

Score interpretation. *Score range.* 0–46 points.

Interpretation of the score. At diagnosis (by definition), the SDI score is 0. Damage is considered if the score is ≥1. Cumulative damage is a poor prognostic sign and a predictor of mortality.

Respondent burden. A complete history and physical is needed. The time-limiting step in completing the instrument is related more to the duration of illness because of the need to review old charts. To complete the form in an essentially well patient with a short history takes <1 minute. For a complex patient not well known to the physician but followed prospectively it can take up to 15 minutes.

Administrative burden. Up to 15 minutes.

Translations/adaptations. English only. The Lupus Damage Index Questionnaire, which is a self-administered version of the SDI, has been validated in Spanish, Portuguese, and French (55,56).

Psychometric Information

Method of development. The SDI was generated by nominal group process. Since the early 1980s, Conference of Prognosis Studies participants were asked to propose a list of items considered to reflect damage in SLE. A list of items that should be included in a damage index, with definitions for ascertainment, was generated. Twenty patient profiles were reviewed by each participant. An item was retained only when there was agreement among the participants that it should be kept in the index.

Reliability. The reliability of the index was shown in a study of 10 "live" patients examined by 6 of 10 physicians from 5 countries representing 10 lupus clinics. The order of patients and physicians was randomized according to a Yonden square design. The SDI detected differences among patients ($P < 0.001$). There was no detectable observer difference ($P = 0.993$) and no order effect ($P = 0.261$) (57).

Validity. *Content and face validity.* In the initial study, 16 of 17 individuals, not members of the SLICC Group, were given the instrument (with suitable instructions). Their scores agreed with the index scores previously determined by the physician who knew the patient's history very well.

Criterion and discriminant validity. Twenty SLICC members completed the index on 42 case scenarios. The intraclass correlation coefficient was 0.553.

Ability to detect change. In a multicenter multiethnic study of 1,297 patients from 8 centers, the SDI showed its ability to record change of damage over time, regardless of the degree of damage recorded for the patients at their first damage index assessment (58).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. This instrument provides an opportunity for clinicians and researchers to assess the accumulated damage in patients with SLE, and it also has been shown in a number of studies to be an excellent tool for prognostic studies.

Caveats and cautions. In patients with a long duration of SLE, the accuracy of the SDI score depends on information available.

Clinical usability. The SDI is useful both as a descriptor for the patient population included in studies, and as an outcome measure for therapeutic trials and studies of prognosis.

Research usability. It is recommended for use in clinical trials, both in stratifying patients and as a component of a responder index.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Griffiths B, Mosca M, Gordon C. Assessment of patients with systemic lupus erythematosus and the use of lupus disease activity indices. Best Pract Res Clin Rheumatol 2005;19:685–708.
2. Gladman DD, Goldsmith CH, Urowitz MB, Bacon P, Bombardier C, Isenberg D, et al. Crosscultural validation and reliability of 3 disease activity indices in systemic lupus erythematosus. J Rheumatol 1992; 19:608–11.
3. Ward MM, Marx AS, Barry NN. Comparison of the validity and sensitivity to change of 5 activity indices in systemic lupus erythematosus. J Rheumatol 2000;27:664–70.
4. Cook RJ, Gladman DD, Pericak D, Urowitz MB. Prediction of short term mortality in systemic lupus erythematosus with term dependent measures of disease activity. J Rheumatol 2000;27:1892–5.
5. Danowski A, Magder L, Petri M. Flares in lupus: Outcome Assessment Trial (FLOAT), a comparison between oral methylprednisolone and intramuscular triamcinolone. J Rheumatol 2006;33:57–60.
6. Ibanez D, Gladman D, Urowitz M. Summarizing disease features over time. II. Variability measures of SLEDAI-2K. J Rheumatol 2007;34:336–40.
7. Swaak AJ, van den Brink HG, Smeenk RJ, Manger K, Kalden JR, Tosi S, et al. Systemic lupus erythematosus: clinical features in patients with a disease duration of over 10 years, first evaluation. Rheumatology (Oxford) 1999;38:953–8.
8. Hay EM, Bacon PA, Gordon C, Isenberg DA, Maddison P, Snaith ML, et al. The BILAG index: a reliable and valid instrument for measuring clinical disease activity in systemic lupus erythematosus. Q J Med 1993;86:447–58.
9. Stoll T, Stucki G, Malik J, Pyke S, Isenberg DA. Further validation of the BILAG disease activity index in patients with systemic lupus erythematosus. Ann Rheum Dis 1996;55:756–60.
10. Isenberg DA, Gordon C. From BILAG to BLIPS: disease activity assessment in lupus past, present and future. Lupus 2000;9:651–4.
11. Rahman P, Gladman DD, Urowitz MB, Hallett D, Tam LS. Early damage as measured by the SLICC/ACR Damage Index is a predictor of mortality in systemic lupus erythematosus. Lupus 2001;10:93–6.
12. Stoll T, Sutcliffe N, Mach J, Klaghofer R, Isenberg DA. Analysis of the relationship between disease activity and damage in patients with systemic lupus erythematosus: a 5-year prospective study. Rheumatology (Oxford) 2004;43:1039–44.
13. Strand V. Lessons learned from clinical trials in SLE. Autoimmun Rev 2007;6:209–14.
14. Strand V, Gladman D, Isenberg D, Petri M, Smolen J, Tugwell P. Outcome measure to be used in clinical trials in systemic lupus erythematosus. J Rheumatol 1999;26:490–7.
15. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang DH, and the Committee on Prognosis Studies in SLE. Derivation of the SLEDAI: a disease activity index for lupus patients. Arthritis Rheum 1992;35: 630–40.
16. Moroni G, Doria A, Mosca M, Alberighi OD, Ferraccioli G, Todesco S, et al. A randomized pilot trial comparing cyclosporine and azathioprine for maintenance therapy in diffuse lupus nephritis over four years. Clin J Am Soc Nephrol 2006;1:925–32.
17. Ong LM, Hooi LS, Lim TO, Goh BL, Ahmad G, Ghazalli R, et al. Randomized controlled trial of pulse intravenous cyclophosphamide versus mycophenolate mofetil in the induction therapy of proliferative lupus nephritis. Nefrology (Carlton) 2005;10:504–10.
18. Furie RA, Petri MA, Wallace DJ, Ginzler EM, Merril JT, Stohl W, et al. Novel evidence-based systemic lupus erythematosus responder index. Arthritis Rheum 2009;61:1143–51.
19. Symmons DP, Coppock JS, Bacon PA, Bresnihan B, Isenberg DA, Maddison P, et al. Development of a computerized index of clinical disease activity in systemic lupus erythematosus. Q J Med 1988;69:927–32.
20. Isenberg DA, Rahman A, Allen E, Farewell V, Akil M, Bruce IN, et al. BILAG 2004: development and initial validation of an updated version of the British Isles Lupus Assessment Group's disease activity index for patients with systemic lupus erythematosus. Rheumatology (Oxford) 2005;44:902–6.
21. Yee CS, Isenberg DA, Prabu A, Sokoll K, Rahman A, Bruce IN, et al. BILAG 2004 index captures systemic lupus erythematosus disease activity better than SLEDAI-2000. Ann Rheum Dis 2008;67:873–6.
22. Isenberg DA, Allen E, Farewell V, D'Cruz D, Alarcon GS, Aranow C, et al. An assessment of disease flare in patients with systemic lupus erythematosus: a comparison of BILAG 2004 and the flare version of SLEDAI. Ann Rheum Dis 2011;70:54–9.
23. Yee CS, Farewell V, Isenberg DA, Prabu A, Sokoll K, Teh LS, et al. Revised British Isles Lupus Assessment Group 2004 index: a reliable tool for assessment of systemic lupus erythematosus activity. Arthritis Rheum 2006;54:3300–5.
24. Yee CS, Farewell V, Isenberg DA, Rahman A, Teh LS, Griffiths B, et al. British Isles Lupus Assessment Group 2004 index is valid for assessment of disease activity in systemic lupus erythematosus. Arthritis Rheum 2007;56:4113–9.
25. Yee CS, Farewell V, Isenberg DA, Griffiths B, Teh LS, Bruce IN, et al. The BILAG 2004 index is sensitive to change for assessment of SLE disease activity. Rheumatology (Oxford) 2009;48:691–5.
26. Yee CS, Cresswell L, Farewell V, Rahman A, Teh LS, Griffiths B, et al. Numerical scoring for the BILAG 2004 index. Rheumatology (Oxford) 2010;49:1665–9.
27. American College of Rheumatology Ad Hoc Committee on Systemic Lupus Erythematosus Response Criteria. The American College of Rheumatology response criteria for systemic lupus erythematosus clinical trials: measures of overall disease activity. Arthritis Rheum 2004; 50:3418–26.
28. Mosca M, Chimenti D, Pratesi F, Baldini C, Anzilotti C, Bombardieri S, et al. Prevalence and clinico-serological correlations of anti-α-Enolasa, anti-C1q, and anti-dsDNA antibodies in patients with systemic lupus erythematosus. J Rheumatol 2006;33:695–7.

29. Amital H, Szekanecz Z, Szucz G, Danko K, Nagy E, Csepany T, et al. Serum concentration of 25-OH vitamin D in patients with systemic lupus erythematosus (SLE) are inversely related to disease activity: is it time to routinely supplement patients with SLE with vitamin D? Ann Rheum Dis 2010;69:1155–7.
30. Vitali C, Bencivelli W, Isenberg DA, Smoler JS, Snaith ML, Scinto M, et al. Disease activity in systemic lupus erythematosus: report of the Consensus Study Group of the European Workshop for Rheumatology Research II. Identification of the variables indicative of disease activity and their use in the developer of an activity score. Clin Exp Rheumatol 1992;10:541–7.
31. Mosca M, Neri R, Giannessi S, Totti D, Bencivelli W, Bombardieri S. The inter-observer variability of the ECLAM index in the evaluation of disease activity in systemic lupus erythematosus [abstract]. Arthritis Rheum 1999;42 Suppl:S98.
32. Mosca M, Bencivelli W, Vitali C, Carrai P, Bombardieri S. The validity of the ECLAM index for the retrospective evaluation of disease activity in systemic lupus erythematosus. Lupus 2000;9:445–50.
33. Fellows of Harvard College. SLE Activity Measure-Revised (SLAM-R). 1988.
34. Chang ER, Abrahamowics M, Ferland D, Fortin PR. Organ manifestations influence differently the responsiveness of 2 lupus disease activity measures, according to patients' or physicians' evaluations of recent lupus activity. J Rheumatol 2002;29:2350–8.
35. Zhang J, Gonzales LA, Roseman JM, Vila LM, Reveille JD, Alarcon GS. Predictors of the rate of change in disease activity over time in LUMINA, a multiethnic US cohort of patients with systemic lupus erythematosus: LUMINA LXX. Lupus 2010;19:727–33.
36. Liang MH, Socher SA, Larsen MG, Schur PH. Reliability and validity of six systems for the clinical assessment of disease activity in systemic lupus erythematosus. Arthritis Rheum 1989;32:1107–18.
37. Bae SC, Koh HK, Chang DK, Kim MH, Park JK, Kim SY. Reliability and validity of Systemic Lupus Activity Measure-Revised (SLAM-R) for measuring clinical disease activity in systemic lupus erythematosus. Lupus 2001;10:405–9.
38. Gladman DD, Goldsmith CH, Urowitz MB, Bacon P, Bombardier C, Isenberg D, et al. Sensitivity to change of 3 systemic lupus erythematosus disease activity indices: international validation. J Rheumatol 1994;21:1468–71.
39. Fortin PR, Abramowicz M, Neville C, du Berger R, Fraenkel L, Clarke AE, et al. Impact of disease activity and cumulative damage on the health of lupus patients. Lupus 1998;7:101–7.
40. Karlson EW, Daltroy LH, Rivest C, Ramsey-Goldman R, Wright EA, Partridge AJ, et al. Validation of a systemic lupus activity questionnaire (SLAQ) for populations studies. Lupus 2003;12:280–6.
41. Trupin L, Tonner MC, Yazdany J, Julian LJ, Criswell LA, Katz PP, et al. The role of neighborhood and individual socioeconomic status in outcomes of systemic lupus erythematosus. J Rheumatol 2008;35:1782–8.
42. Wolfe F, Petri M, Alarcon GS, Goldman J, Chakravaty EF, Katz RS, et al. Fibromyalgia, systemic lupus erythematosus (SLE), and evaluation of SLE activity. J Rheumatol 2009;36:82–8.
43. Yazdany J, Yelin EH, Panopalis P, Trupin L, Julian L, Katz PP. Validation of the Systemic Lupus Erythematosus Activity Questionnaire in a large observational cohort. Arthritis Rheum 2008;59:136–43.
44. Committee on Prognosis Studies. Prognosis studies in SLE: an activity index [abstract]. Arthritis Rheum 1986;29 Suppl:S93.
45. Gladman DD, Ibanez D, Urowitz MB. Systemic Lupus Erythematosus Disease Activity Index 2000. J Rheumatol 2002;29:288–91.
46. Touma Z, Urowitz MB, Gladman DD. SLEDAI-2K for a 30-day window. Lupus 2010;19:49–50.
47. Uribe AG, Vila LM, McGwin G Jr, Sanchez ML, Reveille JD, Alarcon GS. The Systemic Lupus Activity Measure-Revised, the Mexican Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), and a modified SLEDAI-2K are adequate instruments to measure disease activity in systemic lupus erythematosus. J Rheumatol 2004;31:1934–40.
48. Petri M, Kim MY, Kalunian KC, Grossman J, Hahn BH, Sammaritano LR, et al. Combined oral contraceptives in women with systemic lupus erythematosus. N Engl J Med 2005;353:2550–8.
49. Sanchez-Guerrero J, Uribe AG, Jimenez-Santana L, Mestanza-Peralta M, Lara-Reves P, Seuc AH, et al. A trial of contraceptive methods in women with systemic lupus erythematosus. N Engl J Med 2005;353: 2539–49.
50. Petri M, Buyon J, Skovron ML, Kim M, for the SELENA Group. Disease activity and health status (SF-36) in postmenopausal systemic lupus erythematosus: the SELENA trial [abstract]. Arthritis Rheum 1997;40 Suppl:S208.
51. Petri M, Buyon J, Kim M. Classification and definition of major flares in SLE clinical trials. Lupus 1999;8:685–91.
52. Guzman J, Cardiel MH, Arce-Salinas A, Sanchez-Guerrero J, Alarcon-Segovia D. Measurement of disease activity in systemic lupus erythematosus: prospective validation of 3 clinical indices. J Rheumatol 1992;19:1551–8.
53. Gladman D, Ginzler E, Goldsmith C, Fortin P, Liang M, Urowitz M, et al. The development and initial validation of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index for systemic lupus erythematosus. Arthritis Rheum 1996;39:363–9.
54. Stoll T, Seifert B, Isenberg DA. SLICC/ACR Damage Index is valid and renal and pulmonary organ scores are predictors of severe outcome in patients with systemic lupus erythematosus. Br J Rheumatol 1996;35: 248–54.
55. Costenbader KH, Khamashta M, Ruiz-Garcia S, Perez-Rodriguez MT, Petri M, Elliot J, et al. Development and initial validation of a self-assessed lupus organ damage instrument. Arthritis Care Res 2010;62: 559–68.
56. Pons-Estel BA, Sanchez-Guerrero J, Romero-Diaz J, Iglesias-Gamarra A, Bonfa E, Borba EF, et al. Validation of the Spanish, Portuguese and French versions of the Lupus Damage Index questionnaire: data from North and South America, Spain and Portugal. Lupus 2009;18:1033–52.
57. Gladman DD, Urowitz MB, Goldsmith CH, Fortin P, Ginzler E, Gordon C, et al. The reliability of the Systemic Lupus Collaborating Clinics/ American College of Rheumatology Damage Index in patients with systemic lupus erythematosus. Arthritis Rheum 1997;40:809–13.
58. Gladman DD, Goldsmith C, Urowitz MB, Bacon P, Fortin P, Ginzler E, et al. The Systemic Lupus International Collaborating Clinics/ American College of Rheumatology (SLICC/ACR) Damage Index for SLE international comparison. J Rheumatol 2000;27:373–6.

Summary Table for Adult SLE Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
BILAG 2004	Disease activity within last 1 month	Physician completed	5–20 min	5–20 min, plus time to complete history and physical examination	Each manifestation scored 0–4 within each organ/system. All results combined into an activity score rated from A (very active) to E (not or never active)	Excellent	Excellent	Excellent	Only validated lupus activity index that aims to show activity in individual systems rather than combining them into a global score	Formal training needed
ECLAM	Disease activity within last 1 month	Physician completed	5–10 min	5–10 min, plus time to complete history and physical examination	Range 0–17.5 Each item scored from 0.5–2.0	Excellent	Excellent	Excellent	Directly derived from a large multicenter study	Global score will miss changes in severity over time
SLAM-R	Disease activity within last 1 month	Physician completed	10–15 min	10–15 min, plus time to complete history and physical examination	Score range 0–81	Excellent	Excellent	Excellent	Includes both dimensions: disease activity and disease severity	Numbers of items are subjective. Scoring relies on the reporting of symptoms by the patients rather than objective documentations
SLAQ	Disease activity within last 3 months	Patient self-completed	Up to 10 min	Up to 10 min	Score range 0–44 Correlates with the physician-completed SLAM	Excellent	Excellent	Excellent	Unique instrument developed and validated for measuring disease outside the clinical setting	If level of education may influence, the response rate still needs to be evaluated
SLEDAI-2K	Disease activity within last 10 days	Physician completed	5–20 min	5–20 min, plus time to complete history and physical examination	Score range 0–105	Excellent	Excellent	Excellent	All versions are validated and widely used by lupus researchers	Does not record improving or worsening. Does not include severity within an organ system
SDI	Disease damage, present for ≥6 months or irreversible event, i.e., surgery	Physician completed	5–15 min	5–15 min, longer time needed to review medical records in complicated cases	Score range 0–46 Damage is considered if score if >0	Excellent	Good	Good	Excellent tool for prognostic studies	In patients with a long duration of SLE, the accuracy of the index depends on information available

* SLE = systemic lupus erythematosus; BILAG 2004 = updated version of British Isles Lupus Assessment Group; ECLAM = European Consensus Lupus Activity Measurements; SLAM-R = Systemic Lupus Activity Measure, Revised; SLAQ = Systemic Lupus Activity Questionnaire for Population Studies; SLEDAI-2K = Systemic Lupus Erythematosus Disease Activity Index 2000; SDI = Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S454–S466
DOI 10.1002/acr.20556

Measures of Depression and Depressive Symptoms

Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9)

KAREN L. SMARR[1] AND AUTUMN L. KEEFER[2]

INTRODUCTION

This article presents a summary of self-report adult measures that are considered to be most relevant for the assessment of depression in the context of rheumatology clinical and/or research practice. This piece represents an update of the special issue article that appeared in *Arthritis Care & Research* in 2003; the current review followed similar selection criteria for inclusion of assessment tools. Specifically, measures were selected based on several considerations, including ease of administration, interpretation, and adoption by arthritis health professionals from varying backgrounds and training perspectives; self-report measures providing data from the patient or research participant's perspective; availability of adequate psychometric literature and data involving rheumatology populations; and frequent use in both clinical and research practice with adult rheumatology populations. This study was not intended to be exhaustive. Clinician-administered, semistructured depression interviews requiring specialized training such as the Hamilton Rating Scale for Depression and Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition were not included. Additionally, measures without sufficient use within rheumatology populations, such as the World Health Organization Composite International Diagnostic Interview depression module and the National Institutes of Health Patient-Reported Outcomes Measurement Information System, were also not included in this review.

Self-report measures that have been included in this review are as follows: Beck Depression Inventory-II, Center for Epidemiologic Studies Depression Scale, Geriatric Depression Scale, Hospital Anxiety and Depression Scale, and Patient Health Questionnaire-9. Some of these measures have become integrated into routine clinical practice (as screening tools) in large managed-care organizations, and these specifics have been included in this article. Included within each measure review are "additional references" that, while not cited within the review itself, may be of interest to the arthritis health professional who intends to use this instrument in their clinical practice or as part of a research study.

As a general comment regarding any assessment of depression, while care was taken to include measures that require little training to administer and interpret, users without psychological background/experience in the management of clinical issues related to depression and crisis situations may need contingency plans for clinical supervision and/or referral sources. Any individual meeting or close to meeting the diagnostic criteria for depressive disorders needs appropriate management and/or referral, including being provided with referral options for different treatment approaches (pharmacologic and/or psychological). Additionally, suicide risk associated with depression must be taken seriously and promptly addressed. Clinicians should have existing plans to immediately deal with anyone who is an imminent danger to self or others (including mandated reporting). Researchers and clinicians ought to identify behavioral health experts (e.g., psychologists, psychiatrists, social workers) who can assist with appropriately handling these types of crisis situations should they be identified in the context of rheumatology clinical or research environments.

BECK DEPRESSION INVENTORY-II (BDI-II)

Description

Purpose. To measure depression symptoms and severity in persons ages ≥13 years.

The views expressed herein are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States Government.

[1]Karen L. Smarr, PhD: Harry S. Truman Memorial Veterans' Hospital and University of Missouri School of Medicine, Columbia; [2]Autumn L. Keefer, PhD: Harry S. Truman Memorial Veterans' Hospital, Columbia, Missouri.

Address correspondence to Karen L. Smarr, PhD, Harry S. Truman Memorial Veterans' Hospital, 800 Hospital Drive, Research Service, Columbia, MO 65201. E-mail: smarrk@health.missouri.edu.

Submitted for publication December 7, 2010; accepted in revised form July 5, 2011.

Versions. Versions include BDI-I (1), BDI-IA (2), BDI-II (3), and BDI for Primary Care (BDI-PC), now known as BDI FastScreen for Medical Patients (BDI-FS) (4).

The Beck Depression Inventory (BDI) has gone through multiple revisions. The original BDI instrument was developed in 1961 (BDI) (1). Revision began in 1971 to improve wording of items, with the final revised instrument published in 1979 (BDI-IA) (2). A technical manual for the BDI-IA was published in 1987 and revised in 1993 (5). The BDI-IA, which is commonly referred to in the literature as simply "BDI," is similar to the original, except the timeframe extends "over the past week, including today," and some items were reworded to avoid double negative statements. The BDI-II, published in 1996, contains a substantial revision of the original and revised BDI-IA, and omits items relating to weight loss, body image, hypochondria, and working difficulty so that the assessment of symptoms corresponds to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria (3). The BDI-II timeframe extends for 2 weeks to correspond with the DSM-IV criteria for major depressive disorder.

The BDI-FS (formerly known as the BDI-PC) contains 7 cognitive and affective items from the BDI-II to assess depression in individuals with biomedical or substance abuse problems (4). The BDI-FS excludes some of the somatic items from the BDI-II. The timeframe on the BDI-FS is the same as the BDI-II.

Populations. The BDI-IA was developed and validated using psychiatric and normal populations. Beck and Steer (5) studied outpatient samples that included persons with severe psychiatric diagnoses, depressive disorders, and substance abuse problems, and college students. The BDI-II was validated using college students, adult psychiatric outpatients, and adolescent psychiatric outpatients (3).

The BDI-FS was validated using general medical inpatients referred for psychiatric consultation and outpatients seen by family practice, pediatrics, and internal medicine (4).

Developer. Aaron T. Beck, PhD, Center for Cognitive Therapy, Philadelphia, Pennsylvania.

Content. The BDI-II was developed to correspond to DSM-IV criteria for diagnosing depressive disorders and includes items measuring cognitive, affective, somatic, and vegetative symptoms of depression (3).

Number of items. There are 21 items in the BDI-IA and BDI-II, and 7 items in the BDI-FS.

Subscales. None.

Recall period for items. Last 2 weeks.

Endorsements. The BDI-II is 1 of 3 instruments (BDI-II, Hospital Anxiety and Depression Scale, Patient Health Questionnaire-9) endorsed by the National Institute for Health and Clinical Excellence for use in primary care in measuring baseline depression severity and responsiveness to treatment.

Examples of use. The original BDI and subsequent versions have been widely accepted and used in psychology and psychiatry for assessing the intensity of depression in psychiatric and normal populations. Studies have been conducted in a variety of settings using medical populations (e.g., Parkinson's disease, human immunodeficiency virus, oncology, cardiac patients, primary care, chronic pain), persons with disabilities, (e.g., arthritis, spinal cord injury, amputation), medically ill persons of diversity, veterans, students, older adults, adolescents, and many populations with psychiatric diagnoses (e.g., eating disorders, addictions, anxiety disorders).

Practical Application

How to obtain. Contact Pearson Assessments (Pearson Assessments, 19500 Bulverde Road, San Antonio, TX 78259-3701, online at www.pearsonassessments.com) to purchase the BDI-II and BDI-FS manuals and instrument; the BDI is no longer sold to the public. Computer software is available from Pearson Assessments for onscreen administration, or for the input of data from a desktop scanner. The computer program may be used to administer a single questionnaire or to integrate the results of sequential administrations.

Method of administration. Paper and pencil self-report in group or individual format; self or oral administration.

Responses. *Scale.* A 4-point scale indicates degree of severity; items are rated from 0 (not at all) to 3 (extreme form of each symptom).

Score range. BDI-II: 0–63, BDI-FS: 0–21, BDI-IA: 0–63.

Scoring. Sum the severity ratings of each depression item. Use the highest response when an item has >1 severity rating.

Special instructions: BDI-II. For diagnostic purposes, item 16 (sleep pattern changes) and item 18 (appetite changes) contain 7-point ratings to note increases or decreases in behavior.

Special instructions: BDI-IA. If the examinee is consciously trying to lose weight, then item 19 is not added to total score.

Score interpretation. No arbitrary cutoff score for all purposes to classify different degrees of depression.

The following guidelines have been suggested to interpret the BDI-II (3): minimal range = 0–13, mild depression = 14–19, moderate depression = 20–28, and severe depression = 29–63. In post–myocardial infarction (heart attack) patients, the recommended cutoff value for the BDI-II was ≥16, with a sensitivity of 88.2% and a specificity of 92.1% (6). Other cutoffs have been recommended for specific medical populations (i.e., insomnia).

The following guidelines have been suggested to interpret the BDI-FS (4): minimal = 0–3, mild depression = 4–8, moderate depression = 9–12, and severe depression = 13–21.

The following guidelines have been suggested to interpret the revised BDI (BDI-IA) (5): minimal range = 0–9, mild depression = 10–16, moderate depression = 17–29, and severe depression = 30–63.

Respondent burden. *Time to administer/complete.* Self-administration is 5–10 minutes; oral administration is 15 minutes.

Administrative burden. *Training to administer.* Minimal training is required for paraprofessionals or professionals to administer. A clinician needs to interpret the BDI-II score by paying particular attention to items endorsing self-harm or feelings of helplessness, such as suicide ideation (item 9) and hopelessness (item 2).

Equipment needed. Pencil or pen to indicate response.

Time to score. 5 minutes.

Training to score. Minimal training; 5–10 minutes.

Training to interpret. Minimal.

Translations/adaptations. The BDI-II has been translated into several languages, including Arabic, Chinese, Dutch, Finnish, French, German, Icelandic, Italian, Japanese, Persian, Spanish, Swedish, Turkish, and Xhosa.

Psychometric Information

Method of development. The original BDI was based on clinical observations and patient description; the BDI-II contains items that reflect the cognitive, affective, somatic, and vegetative symptoms of depression (1,2). The BDI-II, a revised version of the BDI-IA, was developed to correspond to DSM-IV criteria for diagnosing depressive disorders (3).

Acceptability. Reading levels vary widely in the literature, ranging from being written at the fifth- to sixth-grade reading level (7,8) to "13 years for the reading level of questions and response options" (9).

Reliability. *Internal consistency.* Beck and Steer (5) report that Cronbach's α for the revised BDI normative–psychiatric samples range from 0.79–0.90. These coefficients are consistent with estimates of coefficient α reported in a psychiatric sample (0.81) (10). The BDI-II has a higher internal consistency than the BDI-IA: Cronbach's α was reported as 0.92 for outpatients and 0.93 for college students. Coefficient α for BDI-FS ranged from 0.85–0.89.

Test–retest. In their review of BDI-IA studies, Beck et al (10) reported correlations between pre- and posttests for varying time intervals that range from 0.48–0.86 for psychiatric patients and from 0.60–0.90 for nonpsychiatric patients. For college students, test–retest correlations ranged from 0.64–0.90; BDI-II test–retest (administered 1 week apart) correlation was 0.93.

Validity. *Content.* According to the manual (5), BDI-IA items reflect 6 of 9 DSM-II criteria well. The BDI-II revision improved content validity by rewording and adding items to assess DSM-IV criteria for depression.

Construct. As theorized, the BDI-IA and BDI-II are positively correlated with hopelessness construct in normative samples. In a factor analysis of the BDI-IA responses of patients and nonpatients, Beck and colleagues (10) found that 3 factors (cognitive–affective, performance, and somatic) were consistently identified across diagnostic groups. Factor analysis of the BDI-II yielded 2 factors (somatic–affective and cognitive factors), a result supported in research with medical outpatients (3,11).

Criterion. In psychiatric outpatient clinic samples, the BDI-II and Hamilton Rating Scale for Depression were positively correlated (0.71) (3).

Sensitivity/responsiveness to change. BDI-II has been found to be sensitive to change in depression in cross-cultural studies: a 5-point difference corresponded to a minimally important clinical difference, 10–19 points to a moderate difference, and ≥20 points to a large difference (12).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Time efficient, simplicity in administration and scoring, psychometric properties, predictive ability, many translations available, used with many different populations, and assessment of symptoms and timeframe of measurement correspond to the DSM-IV criteria.

Caveats and cautions. Concerns about overlapping symptoms between medical conditions and depression, although somatic symptoms have been shown to be an important assessment inclusion for depression in elderly medical patients (13); cost; and reading level. The manual suggests cautious use for diagnosis based on scores alone. Recommend health professionals interpret BDI-II scores and provide indicated referrals/interventions.

Clinical usability. High. Has demonstrated utility in identifying depression in medical populations, including rheumatology.

Research usability. High. Strong psychometric properties support use.

Additional references. Arnau RC, Meagher MW, Norris MP, Bramson R. Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. Health Psychol 2001;20:112–9.

Beck AT, Guth D, Steer RA, Ball R. Screening for major depression disorders in medical inpatients with the Beck Depression Inventory for Primary Care. Behav Res Ther 1997;35:785–91.

Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of the Beck Depression Inventories-1A and -II in psychiatric outpatients. J Pers Assess 1996;67:588–97.

Furukawa TA. Assessment of mood: guides for clinicians. J Psychosomatic Res 2010;68:581–9.

Steer RA, Cavalieri TA, Leonard DM, Beck AT. Use of the Beck Depression Inventory for Primary Care to screen for major depression disorders. Gen Hosp Psychiatry 1999; 21:106–11.

CENTER FOR EPIDEMIOLOGIC STUDIES DEPRESSION SCALE (CES-D)

Description

Purpose. To measure the current level of depressive symptoms in a general population.

Versions. The original 20-item version has been shortened to a 10-item version for older adults (14) and to a 5-item version (15). A 9-item version was developed for screening rheumatoid arthritis (RA) patients (16). Multiple other shortened versions ranging from 4–16 items have been developed and used with various populations, although shorter versions tend to increasingly classify patients with multiple chronic health conditions (including RA) as depressed (17). There is also a modified version available for children (CES-D for Children) (18).

Populations. Epidemiology studies using a general population.

Developer/contact information. Lenore Sawyer Radloff, National Institute of Mental Health, Rockville, Maryland.

Content. Items assess perceived mood and level of functioning during the past week. Four factors are represented: depressed affect, positive affect, somatic problems and retarded activity, and interpersonal relationship problems, with an emphasis on depressed affect. CES-D items do not assess the diagnostic criteria of appetite, anhedonia, psychomotor agitation or retardation, guilt, or suicidality.

Number of items. 20 items.

Subscales. None.

Recall period for items. The past week.

Endorsements. None.

Examples of use. Widely used and validated in many populations, including RA, fibromyalgia, and other medical cohorts (stroke, multiple sclerosis, oncology, spinal cord injury, diabetes mellitus); adolescents; women; diverse populations; primary care; elderly; and clinical and psychiatric populations.

Practical Application

How to obtain. The CES-D is available in original article by Radloff (19) and is available online at www.chcr.brown.edu/pcoc/cesdscale.pdf. There is no cost to use the CES-D; it is available in the public domain.

Method of administration. Easily self-administered or administered by interviewer. Can be administered in-person, by written or interview format, by telephone interview, or by mail.

Responses. *Scale.* 4-point scale, where 0 = rarely or none of the time (<1 day), 1 = some or a little of the time (1–2 days), 2 = occasionally or a moderate amount of time (3–4 days), and 3 = most or all of the time (5–7 days).

Score range. The range is 0–60 for the original 20-item version.

Scoring. Easily hand scored. Items are summed to obtain a total score using the 0 (rarely or none of the time) to 3 (most or all of the time) scores for individual items. Four items (4, 8, 12, 16) are worded in a positive direction to reduce a tendency toward response bias; these items are reverse coded.

Score interpretation. A higher score reflects greater symptoms of depression, weighted by frequency of occurrence in the past week. CES-D score ≥16 is typically employed as a cutoff for clinical depression and usually warrants a referral for a more thorough diagnostic evaluation.

While maximizing sensitivity, the cutoff score of ≥16 results in a high percentage of false-positives; therefore, Haringsma et al (20) note an optimal cutoff score for clinically relevant depression as 22 (with sensitivity 84%, specificity 60%, and positive predicted value 77%) with community-dwelling older adults. Turk and Okifuji (21) recommend a cutoff score of 19 for detecting depressive disorder in patients with chronic pain. Blalock et al (22) identified 4 arthritis-related items and suggested a modified scoring approach. Martens et al found little difference between CES-D cutoff scores of 16 and 19 in RA patients, but noted a score of 16 yielded stronger sensitivity and negative predictive value and a score of 19 yielded stronger specificity and positive predictive value; therefore, the authors recommend the cutoff score of 19 as being optimal in most cases (23). Callahan et al (24) and McQuillan et al (25) discussed additional scoring issues in rheumatic disease.

Respondent burden. Time to administer/complete: ~10 minutes.

Administrative burden. *Training to administer.* Minimal.

Equipment needed. When self-administered, a pencil or pen to complete.

Time to score. Less than 10 minutes. Can be scored during administration.

Training to score. Minimal training time to score; ≤10 minutes.

Training to interpret. ≤10 minutes.

Translations/adaptations. Translated into Arabic, Chinese, Dutch, French, German, Greek, Korean, Italian, Japanese, Portuguese, Russian, Spanish, Turkish, and Vietnamese.

Psychometric Information

Method of development. The CES-D was developed for research purposes and is used as a screening tool to identify persons at risk for clinical depression.

Acceptability. Written at the third-grade reading level (9). Microsoft Word 2007 Flesch-Kincaid analysis completed by the authors reveals a reading grade level of 3.3. Due to length, missing data in the 20-item CES-D have been reported as common in studies of elders.

Reliability. *Internal consistency.* High internal consistency. Coefficient α range from 0.85 in the general population to 0.90 in a psychiatric population.

Test–retest. The CES-D measures "current" level of symptoms and is expected to vary over time. In the original sample, test–retest correlations were in the moderate range, falling between 0.45 and 0.70, as expected if the scale is sensitive to current depressive state; stronger test–retest correlations were identified with shorter administration time intervals.

Validity. *Content.* Items were selected from the longer previously used and validated scales considered to be representative of clinical symptoms of depression.

Criterion. The CES-D is widely studied in the literature and deemed an accepted measure of depression. It adequately correlates with other valid self-report depression scales to provide concurrent validity. In the original sample, CES-D correlations with depression measures (e.g., Lubin, Bradburn Negative Affect) ranged from 0.51–0.61; moderate correlation (0.49) was found between CES-D and clinical interview ratings of depression. CES-D scores were moderately correlated with self-esteem (0.58) and state anxiety (0.44) and highly correlated (0.71) with trait anxiety (26).

Sensitivity/responsiveness to change. Sensitive to change since the test–retest changes have been found before and after treatment, as well as before and after a stressful life event. Authors have been unable to locate any clear cutoff scores for measurement of statistically significant change; however, ranges of 13–21 have been provided for detecting 80–90% reliable change (27).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Validated and used with many different populations, many translations available, various length formats available with differing scoring systems, and cost (free to use). Researchers have used shorter versions with various populations to examine alternate CES-D cutoffs and a simplified scoring system (yes/no) (28).

Caveats and cautions. High false-positive rate for clinical depression with a standard cutoff score of 16. Response format in original 20-item instrument can be difficult for some participants to follow and is a contributing reason for the development of shorter versions; clinicians should be aware of this possibility for difficulty when selecting the original instrument over the shorter versions.

Clinical usability. Moderate to high. A CES-D cutoff score of 16 seems appropriate in most populations, especially when the goal is to identify individuals at high risk for major depressive disorder, accepting some false-positives. Slightly lowering the CES-D cutoff may be necessary to identify persons with dysthymic disorder or minor depressive disorder. The CES-D is not intended as a diagnostic tool (29), to discriminate among depression subtypes (major depressive disorder versus dysthymic disorder; bipolar versus unipolar), or to distinguish between a primary or secondary depression (30).

Research usability. Moderate to high. The CES-D has been extensively used and studied, and is considered a reliable valid instrument and a widely recognized research tool. The CES-D can be used to measure change in affective state and is an excellent choice to measure depression symptoms in research studies. The CES-D can be used in diverse settings and has been validated in numerous populations, allowing comparisons across studies.

The high correlation between CES-D measures and trait anxiety indicates that CES-D measures depression as well as anxiety, a conceptually related construct. Based on the validity studies, the CES-D may not be specific for depression, but may be a measure of general distress. Additionally, the instrument does not specifically address suicidal ideation. Therefore, its utility for research studies is dependent on the specific aspects of depression the researcher seeks to measure and his or her need to exclude other possible constructs in this measurement.

Additional references. DeForge BR, Sobal J. Self-report depression scales in the elderly: the relationship between the CES-D and the Zung. Int J Psychiatry Med 1988;18: 325–8.

Escalante A, del Rincon I, Mulrow CD. Symptoms of depression and psychological distress among Hispanics with rheumatoid arthritis. Arthritis Care Res 2000;13: 156–67.

Foley KL, Reed PS, Mutran EJ, DeVellis RF. Measurement adequacy of the CES-D among a sample of older African Americans. Psychiatry Res 2002;109:61–9.

Gerety MB, Williams JW Jr, Mulrow CD, Cornell JE, Kadri AA, Rosenberg J, et al. Performance of case-finding tools for depression in the nursing home: influence of clinical and functional characteristics and selection of optimal threshold scores. J Am Geriatr Soc 1994;42:1103–9.

Roberts RE. Reliability of the CES-D Scale in different ethnic contexts. Psychiatry Res 1980;2:125–43.

GERIATRIC DEPRESSION SCALE (GDS)

Description

Purpose. Developed as a self-rating screening tool to measure depressive symptoms in older adults. Designed to identify depression in the elderly by distinguishing symptoms of depression and dementia.

Versions. The original or "long version" contains 30 items (31,32). A short version (15 items) was developed to decrease fatigue or lack of focus seen in the elderly (33).

Populations. Normal community-dwelling elderly and elders hospitalized for depression.

Developer. Jerome Yesavage, MD, Director, Stanford University/VA/NIA Aging Clinical Research Center, Palo Alto, California.

Content. Items represent characteristics of depression in the elderly in the affective (e.g., sadness, apathy, crying) and cognitive domains (e.g., thoughts of hopelessness, helplessness, guilt, worthlessness). The GDS contains no somatic concerns common in the elderly (i.e., disturbances in energy level, appetite, sleep, sexual interest).

Number of items. 30 items in the original or long form and 15 items in the short form.

Subscales. None.

Recall period for items. Current for long form; past week for short form.

Endorsements. None.

Examples of use. Stiles and McGarrahan (34) reported that the GDS has been used successfully in community samples, psychiatric and medical patients, nursing home residents (cognitively impaired and intact), geriatric samples, and young adults. The GDS in both formats (original and short) has been widely used with elderly medical patients (i.e., primary care, stroke, rheumatology, Parkinson's disease, and cancer) and persons of diversity (i.e., Asians, African Americans, Mexican Americans).

Practical Application

How to obtain. Available from the original article by Yesavage and colleagues (32); English long and short versions, scoring instructions, and versions in many languages are available online at www.Stanford.edu/~yesavage/GDS.html. There is no cost; the GDS is in the public domain.

Method of administration. Designed as paper and pencil, self-administered questionnaire. Oral assistance and/or interview can be utilized; however, it has been suggested that the same format be used for repeated administrations for patients or within subject groups because different administration formats can produce variable results. Oral administration may be advisable in some situations, particularly for individuals who have cognitive impairments.

Responses. *Scale.* Yes or no.

Score range. The range is 0 (no depression) to 30 (severe depression) for long form, and 0 (no depression) to 15 (severe depression) for short form.

Scoring. *Original/long form.* Total score calculated by summing responses that endorse depression. Negatively endorsing items 1, 5, 7, 9, 15, 19, 21, 27, and 29 indicates depression, while positively endorsing the remaining 20 items indicates depression.

Short form. Consistent with the long form, the total score is calculated by summing responses that endorse depression. Negatively endorsing items 1, 5, 7, 11, and 13 indicates depression, while positively endorsing the remaining 10 items indicates depression.

Missing data. According to the above noted web site, prorating scores is permitted. An example provided on the developer's web site is as follows: "If say 3 of 15 items missed, total score is score on 12 completed PLUS 3/15ths of total score to make-up for omitted items, e.g. if they got a 4 on the 12 they completed or 1/3 positive, add 1/3 of the 3 missing or 1 point for a total of 5."

Score interpretation. *Long form.* Higher GDS scores are indicative of more severe depression. Brink et al (31) suggested GDS scores 1–10 be considered normal, while GDS scores ≥11 are indicative of possible depression; using a cutoff score of 14 avoids false-positives. The developer's web site provides the following interpretive guidelines: 0–9 = normal, 10–19 = mild depression, and 20–30 = severe depression.

Short form. The developer's web site reports scores >5 are suggestive of depression and those >10 indicate highly likely depression. Studies involving medical patients propose cutoffs ranging from 5–7 (35–39).

Respondent burden. Time to administer/complete: 5–10 minutes for the original/long version and 2–5 minutes for the short version.

Administrative burden. *Training to administer.* None.

Equipment needed. Pencil or pen to record responses.

Time to score. 2 minutes.

Training to score. Minimal; 5 minutes.

Training to interpret. Minimal; 5 minutes.

Norms available. None.

Translations/adaptations. The GDS has been translated into Arabic, Chinese, Creole, Danish, Dutch, Farsi, French, French Canadian, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Italian, Japanese, Korean, Lithuanian, Malay, Maltese, Norwegian, Portuguese, Romanian, Russian, Serbian, Spanish, Swedish, Thai, Turkish, Vietnamese, and Yiddish, and are available for download via the developer's web site (www.Stanford.edu/~yesavage/GDS.html).

Psychometric Information

Method of development. Items are based on characteristics of depression in the elderly. Brink and colleagues (31) selected 100 items that distinguished between elderly depressed and nondepressed individuals; 30 items were selected for the GDS using an empirical selection procedure. For the short form, the 15 questions that had the highest correlations with original validation studies were chosen from the pool of 30 (33).

Acceptability. The GDS short form is written at a fourth-grade reading level (9). Microsoft Word 2007 Flesch-Kincaid analysis completed by the authors reveals a reading grade level of 4.8 for the short form and 4.9 for the original long form.

Reliability. *Internal consistency.* High: Cronbach's α was 0.94 and split-half reliability was 0.94 (32). Recent review of internal consistencies measured by multiple studies using both oral and written administrations with various populations reported Cronbach's α ranging from 0.69–0.99 (40) for the long form; short form internal consistency has been reported at 0.74–0.86 (39,41).

Test–retest reliability. Correlations (r = 0.84–0.85) at 1–2 weeks retest suggested the GDS scores (both short and long forms) reflect stable individual differences.

Validity. *Content.* Final test items for both forms selected via empirical item selection from items based on characteristics of depression in the elderly.

Criterion. High correlations have been noted between the GDS and other depressive symptom measures. The GDS more consistently differentiates depressed from nondepressed seniors than other depression measures (34). The GDS short form is highly correlated with the original long form (33).

Construct. Yesavage and colleagues (32) validated the original 30-item version using 2 depressive symptoms measures, the Zung Self-Rating Scale for Depression (SDS) and the Hamilton Rating Scale for Depression (HRSD), to compare their ability to classify normal subjects from mild and severe depression. The measures yielded similar results, with normal subjects scoring lower than persons endorsing mild depressive symptoms and those endorsing severe depressive symptoms, and persons with severe symptoms having the highest scores. When compared to a diagnostic classification variable, the GDS and HRSD yielded similar results, while the SDS appeared to discriminate less effectively. Correlation between the GDS and SDS was 0.84; correlation between the GDS and HRSD was 0.83.

Other studies have used depression measures (i.e., Center for Epidemiologic Studies Depression Scale) to examine the GDS convergent validity. Stiles and McGarrahan (34) reported that most studies report correlations ranging from 0.58–0.89. Studies involving young subjects reported lower correlations (range 0.66–0.67).

Divergent. The correlations between the GDS and cognitive screening tests, Mini-Mental State Examination and modified Blessed Test, were low since they intended to measure different constructs.

Sensitivity/responsiveness to change. Sensitivity of the GDS to change was compared to the SDS and HRSD; normal subjects were expected to receive the lowest GDS scores, and persons reporting with severe depressive symptoms to receive the highest scores. When SDS, HRSD, and GDS scores were compared, the 3 severity levels seen in the GDS were also seen in criterion measures.

The GDS short form has been shown to differentiate between depressed and nondepressed elderly primary care patients with a sensitivity of 0.814 and a specificity of 0.754 at a cutoff score of 6 (39), with a recent meta-analysis of primary care patients providing similar data (original/long form GDS sensitivity = 77.4%, specificity = 65.4%; short form GDS sensitivity = 81.3%, specificity = 78.4%) (41).

The GDS has been shown to be sensitive to change reflecting change in the depression of patients with rheumatoid arthritis following depression treatment, with GDS score changes of 6–11 points needed for 80–90% reliable change (29).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Time efficient; simplicity in administration and scoring, robust psychometric properties of both the long and short forms, many translations available, and extensively studied with the elderly population. The GDS has a simple format that accurately and efficiently assesses depressive symptoms in the elderly, from those ages 65 years to those ages ≥85 years.

Caveats and cautions. The GDS appears valid in younger samples, yet may not be the best choice of assessment with a younger sample. A gap remains regarding the validity of the GDS in persons ages ≥85 years. The GDS may also assess "general distress" rather than only depressive symptoms: several items are indicative of both cognitive and somatic symptoms of anxiety. Simple administration and robust psychometric properties have led to the GDS being translated into many languages and cultures, yet studies conducted in other countries/cultures suggest that depressive symptoms are expressed differently in other parts of the world, suggesting cautious use.

GDS long and short forms were highly correlated (0.84) with and sensitive to depressive symptoms in mild to moderate dementia. Debate in the literature concluded that the GDS was effective and reliable in individuals with mild dementia. Stiles and McGarrahan (34) recommend caution when using the GDS with cognitively impaired individuals, and also recommend not using the scale with severely cognitively impaired patients or individuals with compromised insight and accuracy of self-report.

There is no consensus or "gold standard" as to which form to use: Stiles and McGarrahan (34) recommend using long form versus short form since it is more reliable and valid; however, Mitchell et al (42) recommend the short form versus long form given minimal added detection in the longer form over the short, but a 3–4-minute addition of time to appointment length.

Clinical usability. High. Has demonstrated utility in identifying depression in elderly medical patients.

Research usability. High. Strong psychometric properties support use.

Additional references. Alden D, Austin C, Sturgeon R. A correlation between the Geriatric Depression Scale long and short forms. J Gerontol 1989;44:124–5.

Harper RG, Kotik-Harper D, Kirby H. Psychometric assessment of depression in an elderly generally medical population. J Nerv Ment Dis 1990;178:113–9.

Katz PP, Yelin EH. Prevalence and correlates of depressive symptoms among persons with rheumatoid arthritis. J Rheumatol 1993;20:790–6.

Olin JT, Schneider LS, Eaton EM, Zemansky MF, Pollack VE. The Geriatric Depression Scale and the Beck Depression Inventory as screening instruments in an older adult outpatient population. Psychol Assess 1992;4:190–2.

Rule BG, Harvey HZ, Dobbs AR. Reliability of the Geriatric Depression Scale for younger adults. Clin Gerontol 1989;9:37–43.

Sheikh JI, Yesavage JA, Brooks JO 3rd, Freidman L, Gratzinger R, Hill RD, et al. Proposed factor structure of the Geriatric Depression Scale. Int Psychogeriatr 1991;3:23–8.

HOSPITAL ANXIETY AND DEPRESSION SCALE (HADS)

Description

Purpose. To assess anxiety and depressive symptoms in a general medical population.

Versions. Original. No additional versions have been developed.

Populations. General medical outpatients ages 16–65 years.

Developer/contact information. A. S. Zigmond and R. P. Snaith, St. James' University Hospital at Leeds, Leeds, UK.

Content. There are 7 depression items measuring cognitive and emotional aspects of depression, predominantly anhedonia, intermingled with 7 anxiety items that focus on cognitive and emotional aspects of anxiety. Somatic items relating to emotional and physical disorders are excluded.

Number of items. 14 items.

Subscales. Anxiety subscale (HADS-A) and depression subscale (HADS-D).

Endorsements. The HADS is 1 of 3 instruments (Beck Depression Inventory-II [BDI-II], HADS, Patient Health Questionnaire-9 [PHQ-9]) endorsed by the National Institute for Health and Clinical Excellence for use in primary care in measuring baseline depression severity and responsiveness to treatment.

Examples of use. Used extensively, primarily with psychiatric and medical patients, including the following patient populations: cancer, traumatic brain injury, cardiac, stroke, intellectual disabilities, hepatitis, diabetes mellitus, epilepsy, chronic obstructive pulmonary disease, Parkinson's disease, postpartum women, chronic pain, patients with amputations, and spinal cord injury. Used with rheumatology patients (e.g., lupus, arthritis, fibromyalgia, Sjögren's syndrome), as well as the general population, students, nonpatients, and subjects with chronic medical conditions. Herrmann (43) tabulated HADS literature specifying study type, medical specialty, population, and originating country where validated.

Practical Application

How to obtain. Copyrighted and available from: GL Assessment, The Chiswick Centre, 414 Chiswick High Road, London, W4 5TF, UK. Order via web site: http://www.gl-assessment.co.uk/health_and_psychology/resources/hospital_anxiety_scale/hospital_anxiety_scale.asp?css=1.

A test manual (44) accompanies the scale and describes administration, scoring procedures, and psychometrics. Additional scoring forms can be ordered via the web site as well. Items comprising the scale can be viewed in the article by Zigmond and Snaith (45).

Method of administration. Paper and pencil self-administered questionnaire. In cases of illiteracy or poor vision, oral administration may be used.

Responses. *Scale.* The scale is a 4-point Likert scale, ranging from 0–3.

Score range. 0–42 for the total score; 0–21 for the HADS-A and HADS-D.

Scoring. Sum the ratings of 14 items to yield a total score; sum the rating on 7 items on each subscale to yield separate scores for anxiety and depression.

Missing data. The test administrator's web site recommends that the score for a single missing item from a subscale is inferred by using the mean of the remaining 6 items. If >1 item is missing, then the subscale should be judged as invalid.

Score interpretation. Higher scores indicate greater severity. Zigmond and Snaith (45) originally recommend the following cutoff scores for the subscales: 0–7 = considered noncase, 8–10 = considered possible case, and 11–21 = considered probable case, which have been reclassified and relabeled as follows: 0–7 = normal, 8–10 = mild, 11–15 = moderate, and ≥16 = severe (44).

Respondent burden. Time to administer/complete: ≤5 minutes.

Administrative burden. *Training to administer.* None; designed as easy, short, and to be administered in the clinic.

Equipment needed. Pencil or pen to endorse items.

Time to score. 1–2 minutes.

Training to score. Minimal.

Training to interpret. Minimal.

Translations/adaptations. Available in English, as well as all other languages of Western Europe and many of Eastern Europe and Scandinavia, along with some African and Far East languages, including Arabic, Chinese, Danish, Dutch, Finnish, French, German, Hebrew, Hungarian, Italian, Japanese, Korean, Norwegian, Portuguese, Spanish, Swedish, Thai, and Urdu.

Psychometric Information

Method of development. The 8 items for the HADS-D were originally created by the developers based on symptoms of anhedonia; the 8 items for the HADS-A were chosen from the Present State Examination, as well as the developers' personal research on symptoms of anxiety and the Hamilton Anxiety Scale (45). Somatic symptoms and symptoms of severe mental disorder were excluded. Items comprising the subscales were intercorrelated and the weaker of the 2 items on each subscale removed, resulting in two 7-item subscales comprised of statistically significantly interrcorrelated items on each.

Acceptability. The HADS is written at a third-grade reading level (46).

Reliability. *Internal consistency.* Cronbach's α ranges from 0.78–0.93 for the HADS-A and from 0.82–0.90 for the HADS-D (47). Similar coefficient alphas were observed for translated versions.

Test–retest. High test–retest correlations (r = >0.80) were found after ≤2 weeks and gradually decrease as time lapses (2–6 weeks = 0.73–0.76 and >6 weeks = 0.70).

Validity. *Content.* The HADS relies on anhedonia, not on somatic symptoms, and is sensitive to mild distress as it excludes symptoms of severe mental illness. Construction of the HADS-D minimizes the effect of somatic disorders associated with depression.

Concurrent. Correlations with corresponding measures of the same theoretical construct (i.e., anxiety or depression) were adequate. Significantly higher correlations were found between the HADS-D and observer ratings and self-assessments for depression than with observer ratings and self-ratings of anxiety; a similar finding was identified with measures of anxiety and the HADS-A. Compared to commonly used depression and anxiety measures (BDI, PHQ, State-Trait Anxiety Inventory, Symptom Checklist-90-Revised), correlations with the HADS-D and HADS-A ranged between 0.60 (good) and 0.80 (very good) (47,48).

Discriminant. The correlation between the HADS-A and HADS-D averages 0.56 (range 0.49–0.74), with this 2-dimension factor supported in literature review.

Sensitivity/responsiveness to change. Designed to identify probable "cases" of anxiety or depression, the HADS is not a diagnostic tool and is a poor predictor of making a specific diagnosis (49). Average sensitivities and specificities are ≥0.80, similar to other self-rating screening tools (43,47). The tabulation by Bjelland (2001) estimated HADS sensitivity and specificity at optimal cutoffs. Silverstone (49) and Goldberg (50) compared HADS-D scores with standard clinical assessments in medical patients. Sensitivity estimates ranged from 56–100%, and specificity estimates ranged from 73–94%. Positive predictive values ranged from 19–70%. These estimates favorably compare to studies using the BDI/BDI-II, Center for Epidemiologic Studies Depression Scale, and PHQ/PHQ-9. Scores have also been found responsive to pharmacologic and psychotherapeutic interventions (43).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Time efficient, widely used with many different populations, and many translations available. The HADS is a reliable valid method for assessing emotional distress in medical populations. Despite its brevity, the HADS screens for possible anxiety and depressive symptoms are similar to more comprehensive clinical measures.

Caveats and cautions. A recent review of use in rheumatoid arthritis patients found much larger effect sizes when the HADS was used compared to other measures of depression (50).

Clinical usability. High. The HADS can be used in clinical and research settings, and may be particularly useful when studying the cognitive processes associated with depressive symptoms and anxiety, since it is free of physical symptoms, such as insomnia and weight loss.

Research usability. High. The HADS has good psychometric properties, making it a good choice to measure psychological distress, to differentiate the symptoms of depression and anxiety, or to examine the impact of cognition on depression or anxiety (47).

Additional references. Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br J Gen Pract 2008;58:32–6.

Johnston M, Pollard B, Hennessey P. Construct validation of the Hospital Anxiety and Depression Scale with clinical populations. J Psychosom Res 2000;48:579–84.

Mykleton A, Stordal E, Dahl AA. Hospital Anxiety and Depression Scale: factor structure, item analysis and internal consistency in a large population. Br J Psychol 2001; 179:540–4.

Snaith RP. The Hospital and Anxiety Depression Scale. Health Qual Life Outcomes 2003;1:1–4.

Snaith RP, Zigmond AS. The Hospital Anxiety and Depression Scale [letter]. Br Med J (Clin Res Ed) 1986; 292:344.

PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

Description

Purpose. To detect and measure depression and severity in medical populations in clinical settings.

Versions. The PHQ (and subsequent variants, which include the Brief PHQ, PHQ-9, PHQ-8, and PHQ-2) was developed from the historical Primary Care Evaluation of Mental Disorders (PRIME-MD), which was shortened to maximize clinical usefulness by combining the 2 original components into a 3-page (or 4-page, depending on administrator preference) self-administered version called the PRIME-MD Patient Health Questionnaire (PHQ) (51). A 2-page version, the Brief PHQ, has also been developed. The PHQ-9 and the shorter PHQ-2 are the depression modules of the PHQ and currently the most widely used versions in clinical settings. Another variant, the PHQ-8, is used primarily in research studies and includes all items of the PHQ-9 except the ninth item that pertains to self-harm. There are multiple other variants of the PHQ used to measure anxiety, somatic symptoms, or depression–anxiety–somatic combinations, which can be found on Pfizer's web site (http://www.phqscreeners.com/).

Populations. The PHQ was validated using 3,000 primary care patients in 8 different clinics and 3,000 obstetrics/gynecology patients in 7 different clinics (51,52).

Developers/contact information. Robert L. Spitzer, MD, and Janet B. W. Williams, DSW, Biometrics Research Department, New York State Psychiatric Institute and Columbia University, New York, New York, and Kurt Kroenke MD, Regenstrief Institute and Department of Medicine, Indiana University, Indianapolis, Indiana.

Content. The 9 items on the PHQ-9 consist of the 9 criteria on which the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) depressive disorder diagnoses are based (53).

Number of items. 9 items in the PHQ-9 and 2 in the PHQ-2.

Subscales. PHQ-9: none, PHQ-2: none.

Recall period for items. Last 2 weeks.

Endorsements. The Veterans Health Administration uses the PHQ-2 as its screening tool for depression in primary care, with a positive screen (score of ≥3) triggering request for completion of the full PHQ-9 and/or additional evaluation for suicide risk, which has been recommended by the developers. PHQ-9 is 1 of 3 instruments (Beck Depression Inventory-II [BDI-II], Hospital Anxiety and Depression Scale, PHQ-9) endorsed by the National Institute for Health and Clinical Excellence for use in primary care in measuring baseline depression severity and responsiveness to treatment.

Examples of use. Studies utilizing the PHQ-9 have been conducted in a variety of settings using medical populations (e.g., arthritis, fibromyalgia, cancer, cardiac patients, chronic pain, primary care, postpartum women, diabetes mellitus, epilepsy, substance abuse, human immunodeficiency virus), persons with disabilities (e.g., spinal cord injury, cognitive impairment), older adults, college students, adolescents, persons of diversity, and in the nonmedical general population.

Practical Application

How to obtain. PRIME-MD, the parent instrument of the PHQ-9 and PHQ-2, was developed in part from a grant from Pfizer; Pfizer maintains the following web site: http://www.phqscreeners.com/. The site includes liability disclaimers, the PHQ and Generalized Anxiety Disorder screeners (in multiple languages), as well as the instruction manual and relevant bibliography. The PHQ-2 is not specifically listed, but includes the first 2 items of the PHQ-9. These items can also be seen in the PHQ-2 validation study (54). Downloading is free; there is no cost associated with its use or reproduction.

Method of administration. Pencil and paper self-report or interview.

Responses. *Scale.* A 4-point scale indicates degree of severity; items are rated from 0 (not at all) to 3 (nearly every day).

Score range. PHQ-9: 0–27, PHQ-2: 0–6.

Scoring. Sum the severity ratings of each depression item.

Score interpretation. *Severity.* The developers report the following interpretive guidelines for the PHQ-9 as a severity measure: 1–4 = no depression, 5–9 = mild depression, 10–14 = moderate depression, 15–19 = moderately severe depression, and 20–27 = severe depression (53).

Diagnostic. As a diagnostic measure, the developers recommend a diagnosis of major depressive disorder (MDD) be considered if ≥5 of the 9 symptom criteria have been present at least "more than half the days" in the past 2 weeks, and if 1 of the symptoms is depressed mood or anhedonia or the criteria of "thoughts that you would be dead or of hurting yourself in some way" is present at all. Consideration of diagnosis of other depressive disorders is recommended if 2, 3, or 4 of the 9 symptom criteria have been present at least "more than half the days" in the past 2 weeks, and 1 of the symptoms is depressed mood or anhedonia, with the recommendation that a clinical evaluation be the final determination of depressive disorder diagnosis (53).

Respondent burden. Time to administer/complete: PHQ-9: <3 minutes, PHQ-2: <1 minute.

Administrative burden. *Training to administer.* Minimal.

Equipment needed. Pen or pencil to indicate response.

Time to score. Minimal.

Training to score. Minimal.

Training to interpret. Minimal training is required for health professionals who can provide appropriate psychotherapeutic intervention and referrals to diagnosed individuals. Clinical supervision may be needed; interviewers may need to provide individuals meeting the criteria for depressive disorders with treatment approaches (pharmacologic and/or psychological), including referral options.

Norms available. No.

Translations/adaptations. The PHQ-9 has been widely translated into many languages, including Spanish, French, Arabic, German, Czech, Dutch, Russian, German, Hindi, Italian, Japanese, Portuguese, and Polish, among others; full availability can be found on the web site (www.phqscreens.com/overview.aspx).

Psychometric Information

Method of development. The original PRIME-MD instrument is a 2-stage (screening component with followup interview modules [based on positive screening items]) diagnostic instrument designed for primary care physicians in general medical settings to identify persons with mental disorders (55). The followup interview for positive screening for depression is the PRIME-MD-Mood Module. The PRIME-MD-Mood Module was developed to guide the clinician to a criterion-based diagnosis of depressive disorders based on the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised (DSM-III-R), now updated to the DSM-IV. The PRIME-MD 2-stage components were combined into a single 3-page questionnaire (PHQ) that can be self-administered. The PHQ-9 is the 9-item depression module (equivalent to the PRIME-MD-Mood Module) from the full PHQ. The PHQ-2 consists of the first 2 items on the PHQ-9.

Acceptability. One literature source reported the PHQ-9 is written at the eighth-grade reading level (9); however, Microsoft Word Flesch-Kincaid analysis conducted by the authors revealed reading grade levels of 3.5–5.0 for versions of the PHQ-9 freely accessible via the internet.

Reliability. *Internal consistency.* Cronbach's α was reported by developers to be 0.89 and 0.86 in the validation studies of the PHQ-9 (53).

Test–retest. Correlations between patient self-administered results and telephone reassessment within 48 hours ranged from 0.84–0.95 (53,56) and from 0.81–0.96 at 7-day reassessment (57).

Validity. *Content.* Items developed directly from DSM-III-R criteria, now updated to DSM-IV, thereby a diagnostic tool.

Construct. Interviews with mental health providers revealed a positive predictive value ranging from 31% for a PHQ-9 cutoff of 9 to 51% for a cutoff of 15 in a sample with a 7% prevalence of MDD (56). In this same sample, positive predictive values for MDD of 21% for a PHQ-2 cutoff of 2 and 56% for a PHQ-2 cutoff of 5 were found; for any depressive disorder, positive predictive values of 48% for a PHQ-2 cutoff of 2 to 85% for a PHQ-2 cutoff of 5 were found (54).

Criterion. Severity of depression as measured by the PHQ-9 was found to be highly correlated with scores on the BDI in the general population (r = 0.73) (58). Strong associations were also found between the PHQ-9 and 20-item Short Form Health Survey (SF-20) scores, particularly those scales most strongly related to depression (e.g., mental health), as well as with self-reported disability days, clinic visits, and the amount of difficulty self-attributed to symptoms (56). Similarly strong correlations were found between PHQ-2 and SF-20 scores, with the strongest correlation again with mental health (range 0.63–0.70) (57). In addition, test characteristics (sensitivity, specificity, likelihood ratio, and area under the receiver operator curve) were found to be similar for the PHQ-2 in comparison to the Symptom-Driven Diagnostic System for Primary Care, Medical Outcomes Study, Center for Epidemiologic Studies Depression Scale (CES-D), 10-item CES-D, BDI, and 13-item BDI version (59).

Sensitivity/responsiveness to change. Utilizing a decline in PHQ-9 score of ≥5 points as an indicator of significant response to treatment or reduction in depression is recommended (57,60).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Time efficient, strong psychometric properties, widely used with many different populations, sensitive to treatment, can be used for both depressive disorders diagnostic and depression severity purposes, and available in the public domain.

Caveats and cautions. If using the PHQ-2 and scores are >3, developers recommend administration of the full PHQ-9 and assessment by qualified personnel.

Clinical usability. High. Has demonstrated utility in efficiently identifying depressive disorders and quantifying depression severity in the medical populations, including rheumatology populations.

Research usability. High. Strong psychometric properties support use.

Additional references. Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical setting with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. J Gen Intern Med 2007;22:1596–602.

Hancock P, Larner AJ. Clinical utility of Patient Health Questionnaire-9 (PHQ-9) in memory clinics. Int J Psychiatry Clin Pract 2009;13:188–91.

Kroenke K, Spitzer RL, Williams JB, Lowe B. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32:345–59.

Lowe B, Schenkel I, Carney-Doebbeling C, Gobel C. Responsiveness of the PHQ-9 to pharmacological depression treatment. Psychosomatics 2006;47:62–7.

Margaretten M, Yelin E, Imboden J, Graf J, Barton J, Katz P, et al. Predictors of depression in a multiethnic cohort of patients with rheumatoid arthritis. Arthritis Rheum 2009;61:1586–91.

Sleath B, Chewning B, De Vellis BM, Weinberger M, De Vellis RF, Tudor G, et al. Communication about depression during rheumatoid arthritis patient visits. Arthritis Rheum 2008;59:186–91.

AUTHOR CONTRIBUTIONS

Both authors were involved in drafting the article or revising it critically for important intellectual content, and both authors approved the final version to be published.

REFERENCES

1. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71.
2. Beck AT, Rush AJ, Shaw BF, Emery G. Cognitive therapy of depression. New York: Guilford Press; 1979.
3. Beck AT, Steer RA, Brown GK. Beck Depression Inventory: second edition manual. San Antonio (TX): The Psychological Corporation; 1996.
4. Beck AT, Steer RA, Brown GK. BDI: Fast Screen for medical patients manual. San Antonio (TX): The Psychological Corporation; 2000.
5. Beck AT, Steer RA. Manual for the Beck Depression Inventory. San Antonio (TX): The Psychological Corporation; 1987.
6. Huffman JC, Doughty CT, Januzzi JL, Pirl WF, Smith FA, Fricchione GL. Screening for major depression in post-myocardial infarction patients: operating characteristics of the Beck Depression Inventory-II. Int J Psychiatry Med 2010;40:187–97.
7. Conoley CW. Review of the Beck Depression Inventory (revised edition). In: Kramer JJ, Conoley JC, editors. Mental measurements yearbook. 11th ed. Lincoln (NB): University of Nebraska Press; 1987. p. 78–9.
8. Groth-Marnat G. Handbook of psychological assessment. 4th ed. Hoboken (NJ): Wiley; 2003.
9. Nelson CJ, Cho C, Berk AR, Holland J, Roth AJ. Are gold standard depression measures appropriate for use in geriatric cancer patients? A systematic evaluation of self-report depression instruments used with geriatric, cancer, and geriatric cancer samples. J Clin Oncol 2010;28: 348–56.
10. Beck AT, Steer RA, Garbin M. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. Clin Psychol Rev 1988;8:77–100.
11. Hiroe T, Kojima M, Yamamoto I, Nojima S, Kinoshita Y, Hashimoto N, et al. Gradations of clinical severity and sensitivity to change assessed with the Beck Depression Inventory-II in Japanese patients with depression. Psychiatry Res 2005;135:229–35.
12. Viljoen J, Iverson G, Griffiths S, Woodward T. Factor structure of the Beck Depression Inventory-II in a medical outpatient sample. J Clin Psychol Med Settings 2003;10:289–91.
13. Norris MP, Arnau RC, Bramson R, Meagher MW. The efficacy of somatic symptoms in assessing depression in older primary care patients. Clin Gerontol 2003;27:43–57.
14. Irwin M, Artin KH, Oxman M. Screening for depression in the older adult: criterion validity of the 10-item Center for Epidemiological Studies Depression Scale (CES-D). Arch Intern Med 1999;159:1701–4.
15. Shrout PE, Yager TJ. Reliability and validity of screening scales: effect of reducing scale length. J Clin Epidemiol 198;42:69–78.
16. Martens MP, Parker JC, Smarr KL, Hewett JE, Ge B, Slaughter JR, et al. Development of a shortened Center for Epidemiological Studies Depression Scale for assessment of depression in rheumatoid arthritis. Rehabil Psychol 2006;51:135–9.
17. Zauszniewsk JA, Bekhet AK. Depressive symptoms in elderly women with chronic conditions: measurement issues. Aging Ment Health 2009;13:64–72.
18. Fendrich M, Weissman M, Warner V. Screening for depressive disorder in children and adolescents: validating the Center for Epidemiological Studies Depression Scale for Children. Am J Epidemiol 1990; 131:538–51.
19. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1:385–401.
20. Haringsma R, Engels GI, Beekman AT, Spinhoven P. The criterion validity of the Center for Epidemiological Studies Depression Scale (CES-D) in a sample of self-referred elders with depressive symptomatology. Int J Geriatr Psychiatry 2004;19:558–63.
21. Turk DC, Okifuji A. Detecting depression in chronic pain patients: adequacy of self-reports. Behav Res Ther 1994;32:9–16.
22. Blalock SJ, DeVellis RF, Brown GK, Wallston KA. Validity of the Center for Epidemiological Studies Depression Scale in arthritis populations. Arthritis Rheum 1989;32:991–7.
23. Martens MP, Parker JC, Smarr KL, Hewett JE, Slaughter JR, Walker SE. Assessment of depression in rheumatoid arthritis: a modified version of the Center for Epidemiologic Studies Depression Scale. Arthritis Rheum 2003;49:549–55.
24. Callahan LF, Kaplan MR, Pincus T. The Beck Depression Inventory, Center for Epidemiological Studies Depression Scale (CES-D), and General Well-Being Schedule Depression Subscale in rheumatoid arthritis. Arthritis Care Res 1991;4:3–11.
25. McQuillan J, Fifield J, Sheehan TJ, Reisine S, Tennen H, Hesselbrock V, et al. A comparison of self-reports of distress and affective disorder diagnoses in rheumatoid arthritis: a receiver operator characteristic analysis. Arthritis Rheum 2003;49:368–76.
26. Orme JG, Reis J, Herz EJ. Factorial and discriminant validity of the Center for Epidemiological Studies Depression (CES-D) Scale. J Clin Psychol 1986;42:28–33.
27. Martens MP, VanDyke M, Parker JC, Smarr KL, Hewett JE, Hewett JE, et al. Analyzing reliability of change in depression among persons with rheumatoid arthritis. Arthritis Rheum 2005;53:973–8.
28. Kohout FJ, Berkman, LF, Evans DA, Cornoni-Huntley J. Two shorter forms of the CES-D depression symptoms index. J Aging Health 1993; 5:179–93.
29. Radloff LS, Teri L. Use of the Center for Epidemiological Studies-Depression Scale with older adults. Clin Gerontol 1986;5:119–36.
30. Radloff LS, Locke BZ. The Community Mental Health Assessment Survey and the CES-D Scale. In: Weismann MM, Meyers JK, Ross CG, editors. Community survey of psychiatric disorder. New Brunswick (NJ): Rutgers University Press; 1985. p. 177–89.
31. Brink TL, Yesavage JA, Lum O, Heersema P, Adey MB, Rose TL. Screening tests for geriatric depression. Clin Gerontol 1982;1:37–44.
32. Yesavage JA, Brink TL, Rose TL, Lum D, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 1983;17:37–49.
33. Sheikh JI, Yesavage JA. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. Clin Gerontol 1986;5: 165–73.
34. Stiles PG, McGarrahan JF. The Geriatric Depression Scale: a comprehensive review. J Clin Geropsychol 1998;4:89–110.
35. Marc LG, Raue PJ, Bruce ML. Screening performance of the 15-item Geriatric Depression Scale in a diverse elderly home care population. Am J Geriatr Psychiatry 2008;16:914–21.
36. Cullum S, Tucker S, Todd C, Brayne C. Screening for depression in older medical inpatients. Int J Geriatr Psychiatry 2006;21:469–76.
37. Bijl D, van Marwijk HW, Ader HJ, Beekman AT, de Haan M. Test-characteristics of the GDS-15 in screening for major depression in elderly patients in clinical practice. Clin Gerontol 2005;29:1–9.
38. Weintraub D, Oehlberg KA, Katz IR, Stern MB. Test characteristics of the 15-item Geriatric Depression Scale and Hamilton Depression Rating Scale in Parkinson disease. Am J Geriatr Psychiatry 2006;14:169–75.
39. Friedman B, Heisel MJ, Delavan RL. Psychometric properties of the 15-item Geriatric Depression Scale in functionally impaired, cognitively intact, community-dwelling elderly primary care patients. J Am Geriatr Soc 2005;53:1570–6.
40. Lopez MN, Quan NM, Carvajal PM. A psychometric study of the Geriatric Depression Scale. Eur J Psychol Assess 2010;26:55–60.
41. Van Marwijk HW, Wallace P, de Bock GH, Hermans J, Kaptein AA, Mulder JD. Evaluation of the feasibility, reliability, and diagnostic value of shortened versions of the Geriatric Depression Scale. Br J Gen Pract 1995;45:195–9.
42. Mitchell AJ, Bird V, Rizzo M, Meader N. Diagnostic validity and added value of the Geriatric Depression Scale for depression in primary care: a meta-analysis of the GDS-30 and GDS-15. J Affect Disord 2010;125: 10–7.
43. Hermann C. International experiences with Hospital Anxiety and Depression Scale: a review of validation data and clinical results. J Psychosom Res 1997;42:17–41.
44. Snaith RP, Zigmond AS. The Hospital Anxiety and Depression Scale manual. Windsor, Berkshire (UK): Nfer-Nelson; 1994.
45. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361–70.
46. Lowe B, Spitzer RL, Grafe K, Kroenke K, Quenter A, Zipfel S, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. J Affect Dis 2004;78: 131–40.
47. Bjelland I, Dahl AA, Haut TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. J Psychosom Res 2002;52:69–77.

48. Silverstone PH. Poor efficacy of the Hospital and Anxiety Depression Scale in the diagnosis of major depressive disorder in both medical and psychiatric patients. J Psychosom Res 1994;38:441–50.
49. Goldberg D. Identifying psychiatric illness among general medical patients. Br Med J (Clin Res Ed) 1985;29:161–2.
50. Dickens C, McGowen L, Clark-Carter D, Creed F. Depression in rheumatoid arthritis: a systematic review of the literature with meta-analysis. Psychosom Med 2002;64:52–60.
51. Spitzer RL, Kroenke K, Williams JB, and the Patient Health Questionnaire Primary Care Study Group. Validation and utility of a self-report version of PRIME-MD: the PHQ (Patient Health Questionnaire) primary care study. JAMA 1999;282:1737–44.
52. Spitzer RL, Williams JB, Kroenke K. Validity and utility of the Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. Am J Obstet Gynecol 2000;183:759–69.
53. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Int Med 2001;16:606–13.
54. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003:41;1284–92.
55. Spitzer RL, Williams JB, Kroenke K, Linzer M, de Gruy FV 3rd, Hahn SR, et al. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. JAMA 1994;272:1749–56.
56. Pinto-Meza A, Serrano-Blanco A, Penarrubia MT, Blanco E, Haro JM. Assessing depression in primary care with the PHQ-9: can it be carried out over the telephone? J Gen Int Med 2005;20:738–42.
57. Lowe B, Unutzer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. Med Care 2004;42:1194–201.
58. Martin A, Rief W, Klaiberg A, Braehler E. Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. Gen Hosp Psychiatry 2006;28:71–7.
59. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression: two questions are as good as many. J Gen Int Med 1997;12:439–45.
60. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002;32:1–7.

Summary Table for Measures of Depression and Depressive Symptoms*

Scale	Content	Measure outputs	Number of items	Response format	Method of administration	Time for administration	Validated populations	Psychometric properties		
								Reliability	Validity	Responsiveness
BDI-II	Cognitive, affective, somatic, and vegetative symptoms	Total score	21 items	0–3 rating scale	Self or oral	5–10 minutes self; 15 minutes oral	Psychiatric and normal populations	Excellent	Good	Adequate
CES-D	Positive affect, negative affect, somatic problems, activity level, and interpersonal items	Total score	20 items	4-point Likert scale	Self or oral	~10 minutes	General populations, including RA and fibromyalgia	Excellent	Good	Good
GDS	Affective and cognitive symptoms common in elderly	Total score	30 items (original form); 15 items (short form)	Yes/no	Self or oral; use same format for subsequent administrations	5–10 minutes (original form); 2–5 minutes (short form)	Elderly, hospitalized, and community dwellers	Excellent	Good	Good
HADS	Intermingled depression and anxiety items	Depression and anxiety subscales; 7 items each	14 items	4-point Likert scale; 0–3 rating scale	Self or oral	~5 minutes	General medical outpatients	Good	Good	Good
PHQ-9	DSM-IV depressive disorders diagnostic criteria	Total score as severity measure; DSM-IV criteria used for diagnosis	9 items	Yes/no	Self or interview	3 minutes	Medical populations in clinic settings, including arthritis	Good	Good	Good

* BDI-II = Beck Depression Inventory-II; CES-D = Center for Epidemiologic Studies Depression Scale; RA = rheumatoid arthritis; GDS = Geriatric Depression Scale; HADS = Hospital Anxiety and Depression Scale; PHQ-9 = Patient Health Questionnaire-9; DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S158–S173
DOI 10.1002/acr.20542

Measures of Function in Low Back Pain/Disorders

Low Back Pain Rating Scale (LBPRS), Oswestry Disability Index (ODI), Progressive Isoinertial Lifting Evaluation (PILE), Quebec Back Pain Disability Scale (QBPDS), and Roland-Morris Disability Questionnaire (RDQ)

ROB SMEETS,[1] ALBERE KÖKE,[2] CHUNG-WEI LIN,[3] MANUELA FERREIRA,[4] AND CHRISTOPHE DEMOULIN[5]

INTRODUCTION

Treatment of patients with chronic low back pain and its evolving disability primarily tries to improve the patients' levels of activities and participation. Mostly, self-reported questionnaires have been used for clinical as well as research purposes to assess daily functioning (1,2), of which the most commonly used will be discussed below. However, this information may not necessarily reflect the real capacity of a patient's performance. A recent review showed that the correlation of self-reported disability and physical activity level was at best moderate for patients with chronic low back pain (3). In order to improve objectivity, measures of body function, e.g., spinal mobility and lumbar extensor muscle strength, have been used, although the correlation with the level of disability is very weak (4,5). Furthermore, there are major concerns about reliability and validity (6–8).

Besides the self-reported disability measures, many have urged to use more objective and direct measures of low back pain–specific functional capacity (5,9,10). Capacity is defined as the highest probable level of functioning that a person may reach in an activity domain at any given moment in a standardized environment. Although there is still no consensus for the definition of functional capacity evaluation (FCE), in the past decades, several FCE measures have been developed, of which the Isernhagen Work Systems Functional Capacity Evaluation (IWS-FCE) is among the most frequently used (11,12). However, recently published psychometric data have shown that some of the tasks included in the IWS-FCE are not reliable (13,14). Unfortunately, the entire sequence of tasks in, for example, the IWS-FCE, is time consuming and expensive, as is the training of the test observer. Therefore, we have decided not to include these measures in this review.

Nevertheless, in order to keep up with, for example, the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials recommendations to evaluate several core outcome domains, including physical functioning (9), we wanted to include easy to use performance tasks. Several tasks have been described (8,15–17), but most of them are not low back pain specific, and some, such as the Back Performance Scale, show insufficient factor structure, as in this measure the quality of the performance is also scored (1,18). Therefore, we decided only to include a performance task that assesses lifting, an activity that specifically might be hampered by low back pain.

For the selection of the self-reported disability questionnaires and lifting performance tasks, we only selected questionnaires/tests that are low back pain specific and of which all psychometric, including responsiveness, properties have been studied in relevant low back pain populations and published in peer-reviewed journals.

Other criteria for selection were: being available in at least English and for performance task measures, easy to administer, inexpensive, and not time consuming when used in clinical practice.

LOW BACK PAIN RATING SCALE (LBPRS)

Description

Purpose. Developed by Manniche et al in 1985, the LBPRS is constructed to measure the 3 clinical illness components of low back pain: pain (back and leg), disability, and physical impairment (19). The scale has been widely used in randomized clinical trials to monitor out-

[1]Rob Smeets, MD, PhD: Centre of Expertise in Rehabilitation and Audiology, Hoensbroek, and Maastricht University, School of Caphri, Maastricht, Limburg, The Netherlands; [2]Albere Köke, PT, MSc: Centre of Expertise in Rehabilitation and Audiology, Hoensbroek, Limburg, The Netherlands; [3]Chung-Wei Lin, PhD: George Institute for Global Health and The University of Sydney, Sydney, New South Wales, Australia; [4]Manuela Ferreira, PhD: George Institute for Global Health, Sydney, New South Wales, Australia; [5]Christophe Demoulin, PT, PhD: University and University Hospital Centre of Liege, Liege, Belgium.

Address correspondence to Rob Smeets, MD, PhD, Zandbergsweg 111, Hoensbroek, Limburg, The Netherlands, 6432 CC. E-mail: r.smeets@adelante-zorggroep.nl.

Submitted for publication February 14, 2011; accepted in revised form June 21, 2011.

come following therapeutic interventions for low back pain (20–28), including older patients (29).

Content. The scale covers 3 domains: back and leg pain (60 points), disability (30 points), and physical impairment (40 points). The first domain includes six 11-point scales, concerning current pain, worst pain in the last 2 weeks, and average pain in the last 2 weeks for both the leg and lower back. The second domain consists of a disability index with 15 questions that range from quality of sleep, social and occupational participation, daily activities, and emotional status. The last domain includes 4 measures of physical impairment: endurance of back muscles, back mobility, overall mobility, and the use of analgesics (19).

Number of items. The pain domain consists of 6 items, the disability domain consists of 15 items, and the physical impairment domain consists of 4 items, yielding a total of 21 self-reported items and 4 performance-based measures.

Response options/scale. The pain domain comprises six 11-point scales, where 0 = "no pain" and 10 = "the worst imaginable pain." The second domain consists of 15 questions and each question is scored from 0–2, where 0 = "not a problem," 1 = "can be a problem," and 2 = "is a problem." The last domain includes 4 measures of physical impairment, each being scored from 0–10 points (19).

Recall period for items. The pain domain includes questions concerning current status and pain intensity in the past 2 weeks. The other 2 domains concern the patient's current status.

Endorsements. The scale has been recommended for functional pain evaluation by researchers in the field (30,31).

Examples of use. The LBPRS has been widely used in randomized clinical trials, in particular those assessing the efficacy of surgical procedures.

Andersen T, Christensen FB, Egund N, Ernst C, Fruensgaard S, Ostergaard J, et al. The effect of electrical stimulation on lumbar spinal fusion in older patients: a randomized, controlled, multi-center trial. Part 2: fusion rates. Spine (Phila Pa 1976) 2009;34:2248–53 (20).

Andersen T, Christensen FB, Ernst C, Fruensgaard S, Ostergaard J, Andersen JL, et al. The effect of electrical stimulation on lumbar spinal fusion in older patients: a randomized, controlled, multi-center trial. Part 1: functional outcome. Spine (Phila Pa 1976) 2009;34:2241–7 (21).

Andersen T, Christensen FB, Hansen ES, Bunger C. Pain 5 years after instrumented and non-instrumented posterolateral lumbar spinal fusion. Eur Spine J 2003;12:393–9 (22).

Filiz M, Cakmak A, Ozcan E. The effectiveness of exercise programmes after lumbar disc surgery: a randomized controlled study. Clin Rehabil 2005;19:4–11 (23).

Radziszewski KR. Comparative retrospective analysis of pain afflictions in patients with lumbar discopathy receiving conservative or operative therapies. Pol Merkur Lekarski 2006;21:335–40. In Polish (24).

Radziszewski KR. The functional status in patients with discopathy of the lumbar spine receiving only conservative therapy or operative therapy. Wiad Lek 2008;61:23–9. In Polish (25).

Soegaard R, Christensen FB, Christiansen T, Bunger C. Costs and effects in lumbar spinal fusion: a follow-up study in 136 consecutive patients with chronic low back pain. Eur Spine J 2007;16:657–68 (26).

Laursen SO, Fugl IR. Outcome of treatment of chronic low back pain in inpatients: effect of individual physiotherapy including intensive dynamic training in inpatients with chronic low back trouble, evaluated by means of low back pain rating scale. Dan Med Bull 1995;42:290–3 (27).

Andersen T, Christensen FB, Niedermann B, Helmig P, Hoy K, Hansen ES, et al. Impact of instrumentation in lumbar spinal fusion in elderly patients: 71 patients followed for 2-7 years. Acta Orthop 2009;80:445–50 (29).

Christensen FB, Stender Hansen E, Laursen M, Thomsen K, Bunger CE. Long-term functional outcome of pedicle screw instrumentation as a support for posterolateral spinal fusion: randomized clinical study with a 5-year follow-up. Spine (Phila Pa 1976) 2002;27:1269–77 (32).

Christensen FB. Lumbar spinal fusion: outcome in relation to surgical methods, choice of implant and postoperative rehabilitation. Acta Orthop Scand Suppl 2004;75: 2–43 (33).

Christensen FB, Hansen ES, Eiskjaer SP, Hoy K, Helmig P, Neumann P, et al. Circumferential lumbar spinal fusion with brantigan cage versus posterolateral fusion with titanium cotrel-dubousset instrumentation: a prospective, randomized clinical study of 146 patients. Spine (Phila Pa 1976) 2002;27:2674–83 (34).

Christensen FB, Laurberg I, Bunger CE. Importance of the back-cafe concept to rehabilitation after lumbar spinal fusion: a randomized clinical study with a 2-year follow-up. Spine (Phila Pa 1976) 2003;28:2561–9 (35).

Videbaek TS, Christensen FB, Soegaard R, Hansen ES, Hoy K, Helmig P, et al. Circumferential fusion improves outcome in comparison with instrumented posterolateral fusion: long-term results of a randomized clinical trial. Spine (Phila Pa 1976) 2006;31:2875–80 (36).

Practical Application

How to obtain. No cost is involved in obtaining the LBPRS. A copy can be downloaded from online outcome measure databases (https://www.cebp.nl/?NODE=77&SUBNODE=1135).

Method of administration. The scale may be completed by either the patient or the interviewer. A modified version of the questionnaire, omitting back muscle endurance, spinal mobility, and total mobility items, has been developed for mail or phone interviews (19).

Scoring. Each of the 3 domains is scored separately and the total score represents a sum of all 3 domains. The score of the first domain ranges from 0–60 points, the score of the second domain ranges from 0–30 points, and the last domain ranges from 0–40 points (19). Together, the 3 domains form a rank scale, where an asymptomatic person scores 0 and a person with extreme disability scores 130 points. However, it is recommended not to use the total sum score, as subscores provide valuable information and are not subject to weighting bias.

Score interpretation. The 3 domains form a rank scale where an asymptomatic person scores 0 and a person with extreme disability scores 130 points. The sum score is influenced by a weighting bias since 3 answer options exist for physical impairment and disability index items and 11 options are used to indicate pain (31).

Respondent burden. In general, the scale is easy to understand and complete. Some items, for instance, items 14 and 15 (item 14: "If it was a present interest do you think that there are certain jobs which you would not be able to manage because of your back trouble?" and item 15: "Do you think that the low back pain will influence your future?"), may be harder to interpret.

Administrative burden. Approximately 15 minutes are required to complete the LBPRS. No training is necessary.

Translations/adaptations. The scale is available in Danish and English (19), Turkish (23), German (37), and Polish (24,25). The scale has been validated in Danish (19) and culturally adapted into German (37).

Psychometric Information

Method of development. The items of the LBPRS were generated to account for the 3 clinical components of low back pain: pain, disability, and physical impairment. Included items concern the etiology of back pain as well as its impact on a patient's psychological social and work status. The scale was developed by a group of researchers and was primarily devised for use in clinical trials; however, it may also be used in clinical settings.

Acceptability. No data on readability or floor or ceiling effects of the scale are available in the literature.

Reliability. The scale presents high interrater reliability (97.7%) (19). No confidence intervals are provided. No information on the minimum detectable change (MDC) or SEM of the scale is available.

Validity. *Content and face validity.* The 4 back pain, leg pain, disability, and physical impairment components of the scale are marginally correlated and yet conditionally independent, suggesting that the LBPRS is a unidimensional scale (latent variable accounts for 65.9% of the total variation of the scale components).

Construct validity. Construct validity was assessed using the conditional Gaussian distribution, where conditional independence among variables was tested using likelihood ratio tests. Results confirmed conditional independence of the LBPRS and doctor's assessment, given the patient's assessment ($P < 0.00005$), and conditional independence of the LBPRS and patient's assessment, given the doctor's assessment ($P < 0.00005$). This suggests that the LBPRS correlates strongly with both the doctor's global assessment and patient's global assessment (19). The German version of the scale presents a high correlation (0.91, $P < 0.000$) with the Roland-Morris Disability Questionnaire (37).

Ability to detect change. Standardized response means for the disability and pain components of the scale are 0.8 (95% confidence interval [95% CI] 0.4–1.3) and 1.3 (95% CI 1.0–1.6), respectively, for patients with low back pain only, and 0.8 (95% CI 0.3–1.2) and 1.3 (95% CI 0.7–1.9), respectively, for patients with low back and leg pain.

Minimum clinically important change (MCIC) was determined by an optimal cut point analysis using both the raw and percent change scores. For the raw scores, the MCIC for the disability and pain components of the scale in all patients was 17 and 10 points, respectively (38).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The 21 self-reported items of the scale are simple and demonstrate a well-balanced distribution of items across the International Classification of Functioning, Disability and Health components (31). It contains items concerning pain, activity limitation, including work activities and activities of daily life, and physical impairment. The pain domain of the scale is responsive and preferable to the numerical rating scale, as it provides more information on pain dimension at 2 different timeframes (38). The scale has been widely used in clinical research, in particular clinical trials involving postsurgical patients.

Caveats and cautions. The disability domain presents lower responsiveness when compared to the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. The LBPRS lacks information on important psychometric properties, such as the MDC and SEM. No information is available on the responsiveness of the performance-based (i.e., physical impairment) component of the scale.

Item-weighting bias has been suggested due to the discrepancy in score ranges across the 3 domains of the scale, and care should be taken when interpreting the total score.

Clinical usability. Information on its MDC and SEM is lacking, but the scale is quick and simple to use and understand and assesses important aspects of the disease (i.e., pain, disability, and physical impairment).

Research usability. Its use in research has been endorsed by experts in the field (30,31). The scale is simple and has been widely used in clinical research in a variety of ways, including face-to-face interviews, mailed follow-ups, and phone interviews.

OSWESTRY DISABILITY INDEX (ODI)

Description

Purpose. The ODI has been developed to assess pain-related disability in people with acute, subacute, or chronic low back pain. Since it was first published in 1980 (version 1.0) (39), several different versions have been developed, including ODI version 2.0, ODI AAOS (modified by the American Academy of Orthopedic Surgeons), and the ODI Chiropractic Version (40). Version 2.0 is recommended for general use (40,41). The rest of this article refers to ODI version 1.0 or ODI version 2.0.

Content. The ODI covers 1 item on pain and 9 items on activities of daily living (personal care, lifting, walking, sitting, standing, sleeping, sex life, social life, and traveling).

Number of items. 10 items.

Response options/scale. Each item is measured on a 6-point ordinal scale, ranging from the best scenario to the worst scenario. For example, for walking (item 4) the response options range from "pain does not prevent me walking any distance" to "I am in bed most of the time and have to crawl to the toilet."

Recall period for items. Version 1.0 is not specific on a timeframe. Version 2.0 relates to "today."

Endorsements. The ODI has been recommended as a back pain–specific measure of disability by researchers in this field (42).

Examples of use. Brox JI, Sorensen R, Friis A, Nygaard O, Indahl A, Keller A, et al. Randomized clinical trial of lumbar instrumented fusion and cognitive intervention and exercises in patients with chronic low back pain and disc degeneration. Spine (Phila Pa 1976) 2003;28:1913–21 (43).

Carette S, Leclaire R, Marcoux S, Morin F, Blaise GA, St-Pierre A, et al. Epidural corticosteroid injections for sciatica due to herniated nucleus pulposus. N Engl J Med 1997;336:1634–40 (44).

Fritzell P, Hagg O, Wessberg P, Nordwall A. 2001 Volvo Award Winner in Clinical Studies. Lumbar fusion versus nonsurgical treatment for chronic low back pain: a multicenter randomized controlled trial from the Swedish Lumbar Spine Study Group. Spine (Phila Pa 1976) 2001;26: 2521–32 (45).

Malmivaara A, Hakkinen U, Aro T, Heinrichs ML, Koskenniemi L, Kuosma E, et al. The treatment of acute low back pain: bed rest, exercises or ordinary activity. N Engl J Med 1995;332:351–5 (46).

Weinstein JN, Tosteson TD, Lurie JD, Tosteson AN, Hanscom B, Skinner JS, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT). A randomized trial. JAMA 2006;296:2441–50 (47).

Practical Application

How to obtain. No permission or cost is required to use the ODI. Copies of the ODI can be found in published sources (40,41).

Method of administration. The ODI is normally completed by patients using paper and pen. Administration by computer (through MODEMS) or telephone is also possible at PO Box 2354, Des Plaines, IL 60017-2354 (40).

Scoring. For each item, the scoring increases incrementally by 1 with each response option, from 0 (first response option) to 5 (last response option). Missing values are omitted. A percentage is worked out to get the total score.

Score interpretation. The total ODI score ranges from 0 (no disability) to 100 (maximum disability). The original developers of the ODI intended for scores from 0–20 to indicate "minimal disability," 20–40 to indicate "moderate disability," 40–60 to indicate "severe disability," 60–80 to indicate "housebound," and 80–100 to indicate "bedbound" (39).

Respondent burden. The ODI is simple to read and can be completed by the respondent in <5 minutes.

Administrative burden. Scoring takes <1 minute. No training is necessary.

Translations/adaptations. The ODI was originally developed in English, but it has been culturally adapted and is available in a range of languages (40,48), such as German, Mandarin, and Spanish.

The ODI Chiropractic Version was developed for patients with less disability, although this version is not recommended by some authors (40).

Psychometric Information

Method of development. The ODI was developed by clinicians at the Robert Jones and Agnes Hunt Orthopaedic Hospital, Oswestry, Shropshire, UK. It is unclear how the items were generated (39).

Acceptability. The ODI is simple to read. The floor or ceiling effects are unclear. Item 8, sex life, has the option of "if applicable" and is at times omitted. An alternative version replaces item 8 by work/housework.

Reliability. The ODI has high internal consistency (Cronbach's $\alpha = 0.71–0.87$) (40) and test–retest reliability (intraclass correlation coefficient 0.84, 95% confidence interval 0.73–0.91) (49). The standard error of measure has been reported to be between 4 and 6 (49,50). Assessed in a group of patients with back pain presented to physiotherapy, the minimal detectable change is 15–19 (49).

Validity. *Content and face validity.* The ODI has adequate content validity, as it covers activities of daily living that are commonly experienced by patients with back pain. However, it lacks generic activities such as work, leisure, recreation, or sporting activities.

Internal construct validity. The ODI has high internal consistency, with Cronbach's alpha between 0.71 and 0.87 (40,41).

External construct/convergent validity. It correlates with other measures of disability, such as the Roland-Morris Disability Questionnaire (RDQ), and shows moderate correlation with pain scales and the Short Form 36 (40,41).

Ability to detect change. There is evidence that the ODI is responsive in detecting change (area under the receiver operating characteristic curve >0.76) (49,51,52). Based on a literature review and discussion, an international panel has suggested 10 points or a 30% score improvement as the cutoff point for minimal important change (53).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The ODI measures pain-related disability, which is an important element affected in people with back pain and a core outcome in this population (42). It is simple to use and score, and has minimal respondent and administrator burden. The ODI has become one of the most commonly used measures of disability in back pain, along with the RDQ. Compared with the RDQ, the ODI is more sensitive in patients with more persistent severe disability, whereas the RDQ is more sensitive to change in patients with mild to moderate disability (2,40).

Caveats and cautions. The ODI has been administered by telephone; however, the multiple response options mean that face-to-face or computer administration would be the preferred method of administration.

Clinical usability. The ODI has established psychometric properties and is easy to use, and therefore is suitable to be used clinically. It can be used both to assess and monitor outcome.

Research usability. The ODI has established psychometric properties and is easy to use, and therefore is suitable to be used in research as a measure of outcome. The ODI is also frequently used as a comparator when evaluating other measures.

PROGRESSIVE ISOINERTIAL LIFTING EVALUATION (PILE), LUMBAR TEST

Description

Purpose. To quantify frequent lifting capacity based on 3 primary limiting factors of patient capability, i.e., psychological, cardiovascular, and anthropometric, while employing isoinertial lifting characteristics. The PILE provides reasonable limits for subject frequent lifting in industry, as well as the limiting factor in lifting (psychophysical or cardiovascular).

By comparing the measured capacity to normative values in industrial workers, the test is able to predict a subject's capacity to tolerate strenuous lifting throughout a day, but it is not sufficient to disqualify applicants or to predict low back pain (LBP) incidents.

It is also used as an outcome measure to evaluate the effect of treatment in patients with chronic LBP (CLBP). The original test was published in 1998, with an erratum notice in 1990 regarding the scoring (54,55).

One modified version has been described in which a starting weight of 4 kg and an incremental weight of 2 kg irrespective of sex are used (56). However, as no other study used this modification, this test will not be discussed.

Content. The participant is asked to lift a box with handles (36 × 26 × 18 cm, 1.35 kg) with an additional weight 4 times in 20 seconds from the floor to a 75-cm high table and back, starting with a total weight of 3.6 kg for women and 5.85 kg for men. After every completed cycle, the weight is increased by 2.25 kg for women and 4.5 kg for men. The test is stopped when the participant is unable to complete 4 lifting cycles within 20 seconds, decides to quit due to fatigue or excessive discomfort (psychophysical end point), when the heart rate (HR) exceeds 85% of the maximal HR (220 − age; cardiovascular end point), when the maximum weight that can be safely lifted has been reached (55–60% of body weight), or when the assessor does not think it is safe to continue the test (safety end point).

Endorsement. The measure has been recommended for more objective functional evaluation in addition to self-report measures (54,55).

Examples of use. Rainville J, Sobel J, Hartigan C, Monlux G, Bean J. Decreasing disability in chronic back pain through aggressive spine rehabilitation. J Rehabil Res Dev 1997;34:383–93 (57).

Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, Van Der Heijden GJ, Knottnerus JA. Chronic low back pain: physical training, graded activity with problem solving training, or both? The one-year post-treatment results of a randomized controlled trial. Pain 2008;134:263–76 (58).

Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, van der Heijden GJ, van Geel AC, et al. Active rehabilitation for chronic low back pain: cognitive-behavioral, physical, or both? First direct post-treatment results from a randomized controlled trial. BMC Musculoskelet Disord 2006;7:5 (59).

Weiner DK, Rudy TE, Glick RM, Boston JR, Lieber SJ, Morrow LA, et al. Efficacy of percutaneous electrical nerve stimulation for the treatment of chronic low back pain in older adults. J Am Geriatr Soc 2003;51:599–608 (60).

Practical Application

How to obtain. The procedure is described in the original publication and erratum notice (54). An extensive protocol in Dutch is available without costs from R. J. E. M. Smeets, MD, PhD (e-mail: r.smeets@adelante-zorggroep.nl).

Method of administration. The assessor increases weight in a standardized manner every 20 seconds and records the maximum weight lifted, the total number of completed lifting cycles, and the HR after each lifting cycle. The assessor also judges whether continuation of the test is safe.

Equipment needed includes a stopwatch, a table of 75 cm height, a box with handles, a set of 2.25 kg and 4.5 kg weights, an HR monitoring system, and paper and pencil.

Scores. Results are expressed as 1) maximum weight lifted, 2) endurance time to discontinuation of test, 3) final and target HRs, 4) total work (sum of forces multiplied by distance), and 5) work consumption (work/time). In order to correct for over- and underweight and facilitate inter-subject comparisons, the "adjusted weight" (AW; derived from a Young Men's Christian Association height/weight chart) normalizing factor can be used (maximum weight lifted to AW).

For outcome measurement, the weight lifted (expressed as a percentage of AW or ideal weight) (57) or the maximum weight lifted in the last fully completed lifting cycle are most commonly used (61–65). In order to analyze results of men and women together, adjustment for the difference in starting weight and incremental weights between men and women is necessary. Therefore, Smeets et al suggested using the number of completed lifting cycles as the main outcome (66–68).

Score interpretation. A normative database based on 61 male and 31 female mixed blue- and white-collar industrial workers (US) is available (54).

Respondent burden. 5–15 minutes; back pain can temporarily increase due to lifting. It is a safe procedure, and none of the studies reported severe side effects.

Administrative burden. 5–15 minutes; instruction of the patient using a written protocol, attaching an HR monitoring system, preparation of box and starting weight, increasing weight every 20 seconds, recording of maximum weight lifted in the last completed cycle or the total number of completed lifting cycles, and HR after each completed lifting cycling on paper.

Psychometric Information

Method of development. A test was developed to measure dynamic lifting capacity without using anatomic stabilization or control of speed/acceleration variables and mimic daily life lifting.

Acceptability. Seven percent (69) to 11% (59) of the patients with LBP were not able to complete 1 lifting cycle before treatment, which might indicate a floor effect.

Reliability. *CLBP subjects.* Interrater reliability for 21 patients with a mean difference of −0.11 kg maximum weight lifted and limits of agreement (LOA) of −2.33 to 2.11 kg was acceptable. The same study using data of 24 patients studied intrarater reliability and found repeatability (2 − SD of mean change) of 4.0 kg (11% of range) in men and 3.6 (18.5% range) in women (62).

Testing with a 2-day interval in 31 patients showed an intraclass correlation coefficient (ICC) of 0.69 for women and 0.91 for men and a smallest detectable change of 6.2 kg maximum weight lifted for women (±3 cycles) and 7.1 kg for men, with a mean test score of 11.8 kg (±4.35 cycles) and 20.8 kg (±4.2 cycles), respectively (65). It should be noted that the patients were instructed to discontinue the test when experiencing an increase of pain or discomfort.

A study using a 5–9-day interval in 50 patients with CLBP found an ICC of 0.92 (95% confidence interval [95% CI] 0.87−0.96) using the number of completed lifting cycles as the outcome (68). The LOA was 2 cycles, which is 48% of the mean score of 4.27 cycles.

Healthy subjects. Test–retest for maximum weight lifted in 10 healthy industrial workers showed a correlation coefficient of 0.87 (54). Another study of 22 female nurses reported an ICC varying from 0.69−0.71 and LOA expressed as the logarithm of time elapsed at termination of 0.75–1.28 and 0.78−1.33 for a 3- and 14-day interval, respectively (70).

The intrarater and interrater reliability in 11 and 12 healthy subjects, respectively, was moderate to good, with a repeatability of 9.37 kg (25% of range) in men and 1.66 kg (8.6% of range) in women, and an interrater repeatability of 5.61 kg (15% of range) in men and 2.37 (12.2% of range) in women (62).

Validity. *Content and face validity.* Measures at the World Health Organization level of activity and by using 3 different end points provide information about a potential limiting factor. Improvement in scores after treatment was similar for different CLPB groups (postdiscectomy versus nonsurgical) (71).

PILE testing showed a sensitivity of 0.85 and a specificity of 0.65 to discriminate between 27 patients with back or neck pain and 26 healthy persons (61).

Construct validity in CLBP subjects. Pearson's correlation with the Work-Well Systems Functional Capacity Evaluation was good (0.75) (72).

In 1 study, 90 subjects with CLBP were randomly assigned to perform at 60% or 100% of their effort. The ability of the tester to differentiate between the amount of effort the subject was lifting on completion of the test had reasonable specificity (84.1%), but unacceptable sensitivity (65.2%) (73).

Correlations with isokinetic lumbar lifting strength before and after CLBP treatment were low and negative for women (−0.08 to −0.32), and positive, although higher prior to treatment, for men (0.38 to 0.63) (55).

Ljungquist et al assessed the influence of pain behavior during testing, pain intensity, duration of pain, more than 1 pain site, sick leave, physical activity during leisure time, and exertion during the test by using linear regression (63). Age, being a woman, and pain during the previous 4 weeks were significantly negatively associated with the PILE results, whereas neck pain and pain in more than 1 side were significantly positively associated.

The influence of psychosocial factors on the performance was confirmed in several studies, although each study included and controlled for many different factors. Geisser et al, while controlling for demographic, physiologic (body mass index [BMI], pain, metabolic equivalents, max HR, perceived effort), and other psychological variables, showed that activity avoidance is significantly associated with the percentage of maximum predicted weight lifted (74). In another study of this group, depression significantly contributed to PILE performance while controlling for age, sex, site of pain, and pain intensity. Furthermore, the physiologic effect during testing (measured by HR) mediated this relationship between depression and performance on the PILE (75). Smeets et al used a linear model, including age, sex, pain, radiating leg pain, duration of symptoms, maximum oxygen consumption ($\mathrm{V\scriptstyle{O}}_{2\max}$), fear of movement/injury, catastrophizing, and internal control, and showed that besides sex, depression and fear of movement significantly, although not very highly, influence the completed lifting cycles (67).

Construct validity in healthy subjects. In a study of 74 healthy women, a linear regression model, including age, height, BMI, strength and endurance of muscle, cardiovascular endurance, trunk mobility, and coordination ability, explained only 40% of the variance with significant associations for flexion mobility, balance, $\mathrm{V\scriptstyle{O}}_{2\max}$, and body height (76). This study confirms that not only physiologic factors are of influence.

Von Garnier et al showed that treatment improving lifting capacity seems to be mediated by reduction of fear-avoidance beliefs about work in nurses with an LBP episode in the last 2 years but not experiencing acute LBP leading to sick leave (77).

Ability to detect change. Mayer et al showed doubling of lifting capacity after work hardening program for CLBP, but provided no effect sizes (55).

Ljungquist et al used a combination of statistical methods to assess whether the PILE is sensitive to pick up clinically important changes in 3 other outcome measures (general health, disturbing pain, and self-efficacy) (64). They conclude that the PILE lumbar test is not responsive to clinically important change. Unfortunately, no raw data such as effect sizes, etc., are provided.

In a study of 223 CLBP patients (mean score 4.2 cycles) with general perceived effect as the external criteria and a threshold of ≥0.70 for the area under receiving operating characteristic curve (AUC) as the criterion for responsiveness, the PILE appeared to be not responsive (AUC 0.59, 95% CI 0.49−0.69) (66). The same study showed that the minimum clinically important change (MCIC) varied from

1.5 (optimal cutoff point based on AUC, sensitivity 0.71, specificity 0.44) to 3.4 (minimum detectable change).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Safe, inexpensive, and easy to administer even by inexperienced persons (e.g., nurse), psychophysical lifting end point, and unconstrained lifting (no anatomical stabilization or control of speed/acceleration variables) truly reflecting self-selected "real-world" lifting techniques.

Caveats and cautions. Cannot be applied in patients taking rate-limiting cardiac medication. Inability to discriminate the "weak link" anywhere along the biomechanical lifting chain.

Approximately 7–11% of patients with CLBP will not be able to complete a lifting cycle before treatment, leaving much room for improvement.

Clinical usability. The results on reliability, especially the LOA, the lack of responsiveness, and a rather high MCIC, are a major concern. Even though most of these studies used the alternative outcome (amount of completed lifting cycles), we recommend not using the PILE as an outcome measure in the treatment of patients with CLBP.

Research usability. There is sufficient evidence on the construct and content validity as well as moderate predictive validity to use the PILE for research, especially for increasing our insight in the complicated interaction between physical and psychosocial factors on frequent lifting, which is often impaired in patients with disabling CLBP. It is an easy to learn and administer test, cheap, and not highly time consuming, which needs only a limited amount of equipment. Despite a potential temporary increase of pain, the test appears to be safe for patients with CLBP.

QUEBEC BACK PAIN DISABILITY SCALE (QBPDS)

Description

Purpose. Measures the level of functional disability (78). This questionnaire was originally developed to monitor and compare patient progress (78). This questionnaire, developed for ambulatory patients with various disability levels and developed for researchers and clinicians, has since been used in various populations (acute low back pain [LBP] [79], chronic disabling pain [75], sacroiliac joint dysfunction [80], lumbar spinal stenosis [81], undergoing disc surgery [82], and posterior surgical decompression [83]) and settings.

The studies describing the development and the measurement properties of the QBPDS were published by Kopec et al in 1996 and 1995, respectively (78,84). According to the results of these initial development studies, the authors suggested a few changes regarding the scale's format and the wording of some of the items to reach the final version of the questionnaire (84).

Content. Items represent elementary daily activities that patients with back pain might perceive difficult to perform. Items can be classified into 6 domains of activity affected by back pain: bed/rest (items 1–3), sitting/standing (items 4–6), ambulation (items 7–9), movement (items 10–12), bending/stooping (items 13–16), and handling of large/heavy objects (items 17–20) (78).

Number of items. 20 items.

Response options/scale. For each item, a 6-point Likert scale (0–5) to indicate the level of difficulty is used, where 0 = "not difficult at all," 1 = "minimally difficult," 2 = "somewhat difficult," 3 = "fairly difficult," 4 = "very difficult," and 5 = "unable to do." Kopec et al suggested using this scale's format rather than the numerical 11-point scale (0–10) used in the development studies (84).

Recall period for items. Patients are asked to answer the QBPDS according to the difficulty they have to perform the activities the current day ("today").

Endorsements. The QBPDS is included in the few back-specific questionnaires recommended in literature (9,10).

Examples of use. Alschuler KN, Theisen-Goodvich ME, Haig AJ, Geisser ME. A comparison of the relationship between depression, perceived disability, and physical performance in persons with chronic pain. Eur J Pain 2008;12:757–64 (75).

Cusi M, Saunders J, Hungerford B, Wisbey-Roth T, Lucas P, Wilson S. The use of prolotherapy in the sacroiliac joint. Br J Sports Med 2010;44:100–4 (80).

Almeida DB, Prandini MN, Awamura Y, Vitola ML, Simiao MP, Milano JB, et al. Outcome following lumbar disc surgery: the role of fibrosis. Acta Neurochir (Wien) 2008;150:1167–76 (82).

Verbunt JA, Sieben JM, Seelen HA, Vlaeyen JW, Bousema EJ, van der Heijden GJ, et al. Decline in physical activity, disability and pain-related fear in sub-acute low back pain. Eur J Pain 2005;9:417–25 (85).

Sanchez K, Papelard A, Nguyen C, Jousse M, Rannou F, Revel M, et al. Patient-preference disability assessment for disabling chronic low back pain: a cross-sectional survey. Spine (Phila Pa 1976) 2009;34:1052–9 (86).

Reneman MF, Jorritsma W, Schellekens JM, Goeken LN. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. J Occup Rehabil 2002;12: 119–29 (87).

Van den Hout JH, Vlaeyen JW, Heuts PH, Zijlema JH, Wijnen JA. Functional disability in nonspecific low back pain: the role of pain-related fear and problem-solving skills. Int J Behav Med 2001;8:134–48 (88).

Wilhelm F, Fayolle-Minon I, Phaner V, Le-Quang B, Rimaud D, Bethoux F, et al. Sensitivity to change of the Quebec Back Pain Disability Scale and the Dallas Pain Questionnaire. Ann Phys Rehabil Med 2010;53:15–23 (89).

Practical Application

How to obtain. A web site (http://www.tac.vic.gov.au/upload/Quebec-Back-Pain.pdf) provides free access to the questionnaire in English. A copy of the questionnaire is also available in the publication by Fritz and Irrgang (79).

Method of administration. The QBPDS is normally completed by patients using paper and pen. It can also be administered by mail (90) or telephone (78).

Scoring. Items are not weighted and the total score is calculated by adding up the scores of each items. There are no specific instructions in case of item omission. Sometimes scores are given for each domain (89).

Score interpretation. Scores range from 0 (no disability) to 100 (maximal disability).

Respondent burden. Self-administration takes ~5 minutes (84,91). No specific difficulty has been reported regarding item reading or understanding.

Administrative burden. Time to administer and to score the questionnaire is short; training necessity is not reported.

Translations/adaptations. The questionnaire has been translated into French (French-Quebec) (78). The use of the French-Quebec version in patients living in France did not cause major problems (91). The QBPDS has been culturally adapted to Dutch (90), Iranian (92), Brazilian Portuguese (93), Turkish (94,95), and Arab (Maroc) (96).

Psychometric Information

Method of development. Several steps (involving clinicians and patients) have been conducted to develop the QBPDS. Forty-eight items designed to assess limitations in elementary activities by using a numerical 11-point scale (ranging from 0 [not difficult at all] to 10 [extremely difficult]) to measure the level of difficulty were administered to 242 ambulatory patients from various settings who sought care for back pain. Patients were asked additional questions concerning item relevance and clarity. Test–retest, responsiveness, and homogeneity of the item analyses were performed; a statistical method based on item-response theory was applied to evaluate the discriminating ability of each item. Final item selection was guided by the analysis as well as by practical considerations. A major concern was to ensure that all types of physical activities relevant to back pain were represented. Developers also wanted the questionnaire to be highly reliable and discriminative over a wide range of disability levels, while at the same time being practical and acceptable to both patients and clinicians (78). Finally, 20 items representing 6 empirically derived categories of activity were selected (84).

Acceptability. The QBPDS appears acceptable to both patients and clinicians (78,91). Kopec et al reported low item omission (range 0.7–1.8%) (84). A higher rate of incomplete questionnaires (10.8%) was reported for questionnaires administered by mail (90).

According to some patients, a few items lack precision and the choice between response options 0 and 1 and between 4 and 5 is not always easy, and the item "throw a ball" surprised some patients (91). No ceiling or floor effects were reported (49).

Reliability. *Internal consistency.* The development study revealed a high internal consistency (Cronbach's α = 0.96) using the original numerical 11-point scale (84). Similar high internal consistency is confirmed for the 6-point Likert scale (0–5) in other languages (Cronbach's α = >0.90) (90,92,93,95).

Reproducibility. Reproducibility is good: the development studies (numerical 11-point scale) revealed high Pearson's correlation coefficients for all items (78) and an intraclass correlation coefficient (ICC) based on 2 self-administrations (spaced by 1–14 days, median 3.8 days) of 0.92 (84).

Davidson and Keating (49) studied test–retest reliability (6-week interval) with the 6-point Likert scale in 47 patients who were seeking treatment for LBP and who reported no change during the 6 weeks. They reported an ICC, SEM, and minimum detectable change (MDC) of 0.84 (95% confidence interval [95% CI] 0.73–0.91), 8 (95% CI 6–10), and 19 (95% CI 14–24), respectively. A similar study in patients with chronic LBP reported a slightly lower SEM (5.7) and smallest detectable change (15.8) (97).

Recently, Hicks and Manal reported an ICC, SEM, and MDC of 0.94, 4.73, and 11.04, respectively, in community-dwelling patients ages 62 years or older with current LBP (mean test–retest interval of 11 days) (98).

Studies in other languages also revealed good reliability, with an ICC generally ≥0.9 (90,93,95).

The test–retest reliability (4-week interval) appeared lower in a group of patients with acute LBP (0.55) (79).

Validity. *Content and face validity.* Content and face validity was good (99), as the questionnaire contains various domains of activity that were selected by patients and health care providers, and has good measurement properties (78). Due to the poor response rate, developers did not include questions on sexual activities, although it may be important (78).

However, although patients were involved in the development questionnaire, disability of the activities assessed by the QBPDS does not necessarily seem to be the priority (86).

Construct validity. The scale is able to discriminate between groups of patients that are expected to differ in the disability level (84) or self-rated health (98).

Internal construct validity. Kopec et al (78) reported a relatively high degree of interitem correlation (ranging from 0.24–0.87) as well as a very high item-total correlation (range 0.59–0.86). Later, the literature reported interitem correlations lower than 0.80 (suggesting absence of redundancy) (91) and item-total correlation ranging from 0.44–0.83 (90,91).

External construct/convergent validity. The QBPDS correlated strongly with other self-reported functional limitation measures such as the Roland-Morris Disability Questionnaire (RDQ), the Oswestry Disability Index (ODI), and the physical function subscale of the Short Form 36 (r = 0.77, 0.80, and 0.72, respectively) (84). Correlations with pain were weak to moderate (r = 0.54) (84).

Recent studies confirmed moderate to strong associations with other disability questionnaires, with correlation coefficients ranging from 0.6–0.91 (87,90–93,95,98), and moderate to weak correlations with pain (90,92,93,95), direct measure of physical function (75,81,87,95), and psychosocial variables (91,98).

Ability to detect change. The few studies dealing with this measurement property (49,79,84,97,100) appeared heterogeneous in populations, the external criterion used, and statistical methods to estimate minimal important change, resulting in a wide range of values.

Responsiveness. Responsiveness was good and similar to the ODI and RDQ (49); in the original studies the items proved highly sensitive to change (78) and the scale appeared able to detect relatively small changes in the level of disability over time (84). Although the difference in change scores between patients who said they had improved and those who said they had deteriorated was significant, the Norman-Streiner coefficient of sensitivity was low (0.26).

Davidson and Keating reported a standardized response mean (SRM) of 0.49 (49); however, Wilhelm et al reported a high sensitivity to change for the QBPDS (SRM 0.80, effect size 0.62) for the total score as for the score of the 6 specific domains (89).

A recent study (97) focusing on the ability of the QBPDS to detect change in patients with chronic LBP referred for a multidisciplinary treatment performed a receiver operating characteristic analysis, and revealed an area under the curve (AUC) of 0.850 points (versus 0.740 and 0.870 in the studies by Davidson and Keating and Fritz and Irrgang, respectively) (49,79). Based on score change (expressed as the percentage) from baseline, the AUC was 0.856 (97).

Interpretability. The literature reported minimal important change (MIC) values ranging from 8.5–32.9 points in patients with back pain (53). This is mainly due to the heterogeneity of the studies. Recently, a group of experts with a particular focus on primary care proposed considering the MIC for the QBPDS as a decrease of 20 points or 30% relative to the baseline score without taking into account the statistical method used (53). However, they specified that different MICs may be more appropriate for different populations and contexts.

A recent study focusing on the ability of the QBPDS to detect change in patients with chronic LBP referred for a multidisciplinary treatment revealed an optimal cutoff value of 5 points based on a receiver operating characteristic analysis (versus 15 points in the study by Fritz and Irrgang [79]), and 18.1% when the score change was expressed as the percentage from baseline (97). This study confirmed that the baseline score has an impact on the magnitude of the optimal cutoff score (97).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. This questionnaire measures the level of functional disability in daily life, which is essential in patients with LBP. Furthermore, it is short, easy to use, and acceptable (for patients and clinicians); has good clinimetric properties (reliability, face and construct validity, ability to detect changes); and is available in several validated translated versions. Therefore, it belongs to the few back-specific disability questionnaires recommended in literature (2).

Caveats and cautions. Because the authors of the original version suggested a few changes (regarding the scale's format and the wording of some of the items) following the 2 development studies (78,84), one cannot be sure that all clinimetric properties reported in those studies are identical for the newly proposed version.

Despite the rather good clinimetric properties, the use of the QBPDS still remains much less frequent than the RDQ or ODI.

Clinical usability. The administrative and respondent burden of the QBPDS is extremely low, and thus easy for clinical use. The absence of a consensus regarding interpretability values resulting from the limited studies with a high level of heterogeneity makes the interpretation of individual score change difficult.

Research usability. The good clinimetric properties of the QBPDS support using it in research.

ROLAND-MORRIS DISABILITY QUESTIONNAIRE (RDQ)

Description

Purpose. The RDQ was designed in 1983 (101) for use in primary care research to assess physical disability due to low back pain (LBP). It has extensively been used in clinical practice in different settings (primary care, injured workers, and multidisciplinary rehabilitation center) to monitor progress in patients with acute, subacute, and chronic LBP and sciatica (1,102). The original description of the RDQ included a pain rating scale that is not recommended and consequently used anymore (41).

Some modifications have been proposed: 1) changing the phrase "because of my back pain" into "because of my back or leg problems" to make it suitable for patients with sciatica (103), 2) reducing the 24 items to 18 due to analysis of redundancy (104), 3) removing 5 items to improve responsiveness and adding 4 additional items, resulting in a 23-item RDQ (103), and 4) changing the timeframe from the last 24 hours into "how many days of the previous month" the patient has been affected (105).

As these modifications resulted in only modest improvements, or have been insufficiently validated, the use of the original version has been recommended (41,42,102). In this review, only data regarding the 24-item original RDQ are shown.

Content. The items represent the execution of daily physical activities and functions that may be affected by LBP, such as housework, sleeping, mobility, dressing, getting help, appetite, irritability, and pain severity. Although it is called a "disability" scale, it contains elements of impairment, disability, and handicap according to the International Classification of Functioning, Disability and Health (78,106,107).

Number of items. 24 items, no subscales.

Response options/scale. In the original version, a patient has to tick a box against the statements that apply to him and leave them blank otherwise. Modified versions use a "yes" and "no" response option for each item (108,109).

Recall period for items. Relates to the last 24 hours.

Endorsements. Deyo et al (42) recommended using the RDQ (or the Oswestry Disability Index [ODI]) in a standard set of outcome measures for back pain. However, since then, in several reviews about functional status measures in back pain, no specific recommendations for a specific measurement tool were made (1,102).

Examples of use. Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, Van Der Heijden GJ, Knottnerus JA. Chronic low back pain: physical training, graded activity with problem solving training, or both? The one-year post-treatment results of a randomized controlled trial. Pain 2008; 134:263–76 (58).

Artus M, van der Windt DA, Jordan KP, Hay EM. Low back pain symptoms show a similar pattern of improvement following a wide range of primary care treatments: a systematic review of randomized clinical trials. Rheumatology (Oxford) 2010;49:2346–56 (110).

Wilkens P, Scheel IB, Grundnes O, Hellum C, Storheim K. Effect of glucosamine on pain-related disability in patients with chronic low back pain and degenerative lumbar osteoarthritis: a randomized controlled trial. JAMA 2010; 304:45–52 (111).

Lamb SE, Hansen Z, Lall R, Castelnuovo E, Withers EJ, Nichols V, et al. Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis. Lancet 2010;375: 916–23 (112).

Hay EM, Mullis R, Lewis M, Vohora K, Main CJ, Watson P, et al. Comparison of physical treatments versus a brief pain-management programme for back pain in primary care: a randomised clinical trial in physiotherapy practice. Lancet 2005;365:2024–30 (113).

Mannion AF, Muntener M, Taimela S, Dvorak J. A randomized clinical trial of three active therapies for chronic low back pain. Spine (Phila Pa 1976) 1999;24:2435–48 (114).

Practical Application

How to obtain. A free download of the questionnaire in different languages is available from www.rmdq.org/Download.htm. Queries may be sent to mroland@man.ac.uk. A copy is available in the original publication (41).

Method of administration. A self-completed questionnaire on paper and an electronic version. Both versions seem equivalent and can be used interchangeably (115). The RDQ can also be administered by telephone (41,109).

Scoring. Items are not weighted. The total score is calculated by adding up the "yes" answers or the items checked by the patient. Scoring does not include an abstinence option (as a result, the denominator remains 24 even if the statement is not applicable to the patient), which may be problematic.

Score interpretation. Scores range from 0 (no disability) to 24 (maximal disability). Female patients more frequently select item 5, "using a handrail to get upstairs," and item 7, "holding on to something to get out of chair," and patients ages ≥65 years more often select item 5 (116,117). In the original study describing the natural history/evolution of LBP, median RDQ scores were 11, 8, and 4 on presentation, 7 days later, and 1 month later, respectively. Stratford et al (118) reported that 68% of patients with mechanical LBP had initial scores ranging from 7–17.

Respondent burden. Completion takes ~5 minutes (119). The RDQ is short and readily understood by patients (52).

Administrative burden. Scoring takes <1 minute. Training is not necessary.

Translations/adaptations. Translations are available in Arabic (Egyptian), Bulgarian, Chinese, Croatian, Czech, Danish (120), Dutch (121), English (Canadian, US, Australian), Flemish, French (122), German (123), Greek (124), Hungarian, Icelandic, Iranian (92), Italian (125), Japanese (126), Korean, Norwegian (127), Polish, Portuguese (128), Brazilian Portuguese (129), Moroccan (130), Romanian, Russian, Spanish (131), Argentinean (132), Columbian, Mexican, Puerto Rican, Venezuelan, Swedish (133), Thai, Tunisian (134), and Turkish (117), as well as for India (Hindi, Kannada, Marathi, Tamil, Telugu, Urdu). Several of these versions have not been validated.

Psychometric Information

Method of development. The RDQ (101) includes 23 items selected from the Sickness Impact Profile (a 136-item health status measure) (135); 15 relate to the physical category, 3 to sleep and rest, 2 to psychosocial, 2 to home management, and 1 to eating. An additional item ("my back is painful almost all the time") is related to the frequency of back pain. The authors chose these items because they describe activities usually affected by LBP; the phrase "because of my back pain" was added to all items to make it specific to LBP and to exclude disability due to another cause.

Acceptability. The RDQ appears acceptable to both patients and clinicians. It is more discriminative in patients who have relatively little disability rather than a high level of disability (41). Proportions of items omitted by the patients are scarcely reported. Kovacs et al (131) and Scharovsky et al (132) reported no missing values compared to 19% and 18%, respectively, in the ODI. The Brazilian RDQ proved to be easy to understand and in 94% of the patients, no item was missing (129). In workers with back injury claims, 14.6% did not answer ≥1 items (109).

Neither floor nor ceiling effects were seen at baseline among workers with recent work-related back injuries (109,136). In a study of patients with mild to moderate low back pain, 22% scored ≤2 at baseline, including 4.9% who scored 0 (137).

Reliability. Good internal consistency is reported, with Cronbach's alpha ranging from 0.84–0.96 (41,106,109,127, 130,138,139). Reliability for short time intervals (1–14 days) (101,132) is higher compared to intervals longer than 6 weeks (49,140). Pearson's correlation coefficients for test–retest in patients with acute/subacute LBP are 0.91 for the same day (101), 0.88 for 1 week (133), and 0.83 for 3 weeks (141). In patients with chronic LBP, a correlation coefficient of 0.72 (interval 2 days to 6 months) was found (140).

The intraclass correlation coefficients (ICCs) for test–retest in patients with acute/subacute LBP are 0.93 for 1–14 days (106), 0.91 for 2 weeks (142), and 0.86 for 3–6 weeks (118). In a mixed group of patients with acute/subacute and chronic LBP referred for physiotherapy, the ICC ranged from 0.42–0.53 (interval of 6 weeks) (49). Almost all studies with a time interval of >2 weeks have lower ICCs than the studies with a shorter interval (142).

In a mixed group of patients with acute/subacute and chronic LBP, SEMs of 3.7 and 4.1, respectively, are reported (49). The SEM depends on the statistical method used, time interval, and definition of unchanged patient, and for patients with chronic LBP it ranged from 1–2.1, 1.3–2.5, and 1.7–2.2, respectively (143).

Minimum detectable change (MDC) also depends on time interval (range 3.7–6.9), definition of "unchanged" patients (range 4.8–6), type of SEM measurement (range 2.7–5.8), treatment type (range 5.4–5.6), and baseline scores (range 5.5–6) (143). Other reports of MDC are generally in line with these ranges (118,127,144,145).

Limits of agreement (LOA) tend to increase as time between tests increases for patients with chronic LBP. Demoulin et al (143) reported almost double values (range −5.8 to 7.8) for a time interval of 12 or more weeks compared to 1–2 weeks (range −3.5 to 3.9). For short intervals (2 weeks), LOA varied from −4.6 to 6.2 (142).

Validity. *Content and face validity.* Only a limited range of problems in physical daily activities related to back pain is assessed. Evaluation of the different activities specified by patients with LBP produced a list of 325 activities (pooled in 56 similar activity groups) compared to the 24 items in the RDQ (146).

The RDQ contains a small number of psychosocial items that are not related to functional limitation per se, e.g., appetite, irritability.

Construct validity. RDQ scores correlate moderately to strongly with other self-reported disability measures: the Quebec Back Pain Disability Scale (r = 0.60) (84,87), the ODI (r = 0.50) (52,87,147), the Back Pain Functional Scale (r = 0.79) (148), the Aberdeen Back Pain Scale (r = 0.68) (149), the Isernhagen Works Systems Functional capacity (r = −0.20) (87), and the EuroQol (r = −0.50) (149,150).

RDQ scores show weak to modest correlations with pain intensity (range 0.26–0.57) (149–151), physical impairment tests such as the straight-leg raising test, and flexion range of motion (range 0.27–0.44) (152).

The RDQ largely satisfies the Rasch model for unidimensionality (108). However, there are insufficient items of higher difficulty to sufficiently evaluate persons with mild disability. Some misfitting items have been found and many of the items are of moderate difficulty with few easy or difficult items. This means that it is easier to detect change for individuals who start with scores in the middle of the range than those who start with high or low scores.

Ability to detect change. The magnitude of responsiveness is dependent on the type of external criteria used (153). Furthermore, the time interval between tests, interpretation of the general perceived effect scale, and baseline scores have a considerable impact on the responsiveness indicators of the RDQ (143).

Several authors comparing the RDQ and ODI have concluded that the RDQ is more sensitive to change (2,102), especially for minor levels of functional limitation. However, the RDQ may be relatively insensitive to deterioration in the patients' condition.

Responsiveness statistics, such as areas under curves (AUCs), ranged from 0.68–0.93 (49,51,52,102,137,154). Demoulin et al (143) reported different AUC scores for different definitions of unchanged patients (range 0.83–0.90) and baseline scores (range 0.89–0.91). Effect sizes ranged from 0.50–1.60 (49,103,106,118,146).

Standardized response means (SRMs) ranged in patients with subacute/chronic LBP from 0.55–0.90 for a time interval of 6 weeks (49,149) to 0.72 (6 months) and 0.83 (1 year) (149). SRMs (3-week period) were 1.34 for patients with acute LBP, 0.80 for patients with subacute LBP, and 0.48 for patients with chronic LBP (155). For another population with chronic LBP (interval of 28 weeks), the SRM ranged from 1.33–2.64 (153).

Cutoff points for relevant improvements strongly depend on baseline severity and methods used for estimation of minimum clinically important change (MCIC) (143,145,153). Systematic reviews concluded that as an approximate guide, changes of 2–3 points on the RDQ between groups should be considered the MCIC (41,145). Kovacs et al (145) reported for patients with subacute and chronic LBP a MCIC ranging from 2.5–6.8 points in patients with baseline scores below 10 points, and from 5.5–13.8 in patients with baseline scores ≥15 points.

Based on an expert consensus, a 30% change from baseline was proposed as a clinically meaningful improvement, which normally means an absolute change of 5 points (53).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The RDQ is the most comprehensively validated measure in low back pain. It is short, simple to complete, and readily understood by patients and clinicians. Psychometric properties are acceptable to good and the RDQ is available in many language versions. It can be used in patients with acute, subacute, and chronic LBP.

Caveats and cautions. There is some evidence that the RDQ does not provide a sufficient spread of items representing activities on a continuum from easy to hard (116). The poor fit of some items to the factor "disability" needs further attention (108,116). Garrat (108) stated that the RDQ could be improved through the removal of items with poor fit statistics and the addition of items toward the extremes of the scale hierarchy. None of the versions have sufficient items of higher difficulty to assess persons with low levels of disability, making it inadequate for assessing function in patients with little disability (116).

Clinical usability. The administrative and respondent burden is very low. RDQ scores and changes scores must be interpreted with caution due to poor-fitting items and the fact that the RDQ does not appear to have interval-level properties. It is inadequate for use in patients with little disability.

Research usability. The psychometric quality is sufficient for using the RDQ in research. Score distributions must be examined before statistical analysis and Rasch-transformed scores can be used to adjust for the imperfections in the scale hierarchy (108).

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Cleland J, Gillani R, Bienen EJ, Sadosky A. Assessing dimensionality and responsiveness of outcomes measures for patients with low back pain. Pain Pract 2011;11:57–69.
2. Grotle M, Brox JI, Vollestad NK. Functional status and disability questionnaires: what do they assess? A systematic review of back-specific outcome questionnaires. Spine (Phila Pa 1976) 2005;30:130–40.
3. Lin CW, Mcauley JH, Macedo L, Barnett DC, Smeets RJ, Verbunt JA. Relationship between physical activity and disability in low back pain: a systematic review and meta-analysis. Pain 2011;152:607–13.
4. Parks KA, Crichton KS, Goldford RJ, McGill SM. A comparison of lumbar range of motion and functional ability scores in patients with low back pain: assessment for range of motion validity. Spine (Phila Pa 1976) 2003;28:380–4.
5. Wittink H. Functional capacity testing in patients with chronic pain. Clin J Pain 2005;21:197–9.
6. Kankaanpaa M, Taimela S, Laaksonen D, Hanninen O, Airaksinen O. Back and hip extensor fatigability in chronic low back pain patients and controls. Arch Phys Med Rehabil 1998;79:412–7.
7. Keller A, Hellesnes J, Brox JI. Reliability of the isokinetic trunk extensor test, Biering-Sorensen test, and Astrand bicycle test: assessment of intraclass correlation coefficient and critical difference in patients with chronic low back pain and healthy individuals. Spine (Phila Pa 1976) 2001;26:771–7.
8. Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine (Phila Pa 1976) 1998;23:2412–21.
9. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9–19.
10. World Health Organization. The International Classification of Functioning, Disability and Health (ICF). 2nd ed. Geneva: World Health Organization; 2001.
11. Soer R, van der Schans CP, Groothoff JW, Geertzen JH, Reneman MF. Towards consensus in operational definitions in functional capacity evaluation: a Delphi survey. J Occup Rehabil 2008;18:389–400.
12. Van Abbema R, Lakke SE, Reneman MF, van der Schans CP, van Haastert CJ, Geertzen JH, et al. Factors associated with functional capacity test results in patients with non-specific chronic low back pain: a systematic review. J Occup Rehabil 2011. E-pub ahead of print.
13. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic low back pain. J Occup Rehabil 2003;13:207–18.
14. Reneman MF, Brouwer S, Meinema A, Dijkstra PU, Geertzen JH, Groothoff JW. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in healthy adults. J Occup Rehabil 2004;14:295–305.
15. Harding VR, Williams AC, Richardson PH, Nicholas MK, Jackson JL, Richardson IH, et al. The development of a battery of measures for assessing physical functioning of chronic pain patients. Pain 1994;58:367–75.
16. Magnussen L, Strand LI, Lygren H. Reliability and validity of the back performance scale: observing activity limitation in patients with back pain. Spine (Phila Pa 1976) 2004;29:903–7.
17. Smeets RJ, Hijdra HJ, Kester AD, Hitters MW, Knottnerus JA. The usability of six physical performance tasks in a rehabilitation population with chronic low back pain. Clin Rehabil 2006;20:989–97.
18. Strand LI, Moe-Nilssen R, Ljunggren AE. Back performance scale for the assessment of mobility-related activities in people with back pain. Phys Ther 2002;82:1213–23.
19. Manniche C, Asmussen K, Lauritsen B, Vinterberg H, Kreiner S, Jordan A. Low Back Pain Rating Scale: validation of a tool for assessment of low back pain. Pain 1994;57:317–26.
20. Andersen T, Christensen FB, Egund N, Ernst C, Fruensgaard S, Ostergaard J, et al. The effect of electrical stimulation on lumbar spinal fusion in older patients: a randomized, controlled, multi-center trial. Part 2: fusion rates. Spine (Phila Pa 1976) 2009;34:2248–53.
21. Andersen T, Christensen FB, Ernst C, Fruensgaard S, Ostergaard J, Andersen JL, et al. The effect of electrical stimulation on lumbar spinal fusion in older patients: a randomized, controlled, multi-center trial. Part 1: functional outcome. Spine (Phila Pa 1976) 2009;34:2241–7.
22. Andersen T, Christensen FB, Hansen ES, Bunger C. Pain 5 years after instrumented and non-instrumented posterolateral lumbar spinal fusion. Eur Spine J 2003;12:393–9.
23. Filiz M, Cakmak A, Ozcan E. The effectiveness of exercise programmes after lumbar disc surgery: a randomized controlled study. Clin Rehabil 2005;19:4–11.
24. Radziszewski KR. Comparative retrospective analysis of pain afflictions in patients with lumbar discopathy receiving conservative or operative therapies. Pol Merkur Lekarski 2006;21:335–40. In Polish.
25. Radziszewski KR. The functional status in patients with discopathy of the lumbar spine receiving only conservative therapy or operative therapy. Wiad Lek 2008;61:23–9. In Polish.
26. Soegaard R, Christensen FB, Christiansen T, Bunger C. Costs and effects in lumbar spinal fusion: a follow-up study in 136 consecutive patients with chronic low back pain. Eur Spine J 2007;16:657–68.
27. Laursen SO, Fugl IR. Outcome of treatment of chronic low back pain in inpatients: effect of individual physiotherapy including intensive dynamic training in inpatients with chronic low back trouble, evaluated by means of low back pain rating scale. Dan Med Bull 1995;42:290–3.
28. Hartvigsen J, Morso L, Bendix T, Manniche C. Supervised and non-supervised Nordic walking in the treatment of chronic low back pain: a single blind randomized clinical trial. BMC Musculoskelet Disord 2010;11:30.
29. Andersen T, Christensen FB, Niedermann B, Helmig P, Hoy K, Hansen ES, et al. Impact of instrumentation in lumbar spinal fusion in elderly patients: 71 patients followed for 2-7 years. Acta Orthop 2009;80:445–50.
30. Longo U, Loppini M, Dnaro L, Maffulli N, Denaro V. Rating scales for low back pain. Br Med Bull 2010;94:81–144.
31. Muller U, Duetz MS, Roeder C, Greenough CG. Condition-specific outcome measures for low back pain part I: validation. Eur Spine J 2004;13:301–13.
32. Christensen FB, Stender Hansen E, Laursen M, Thomsen K, Bunger CE. Long-term functional outcome of pedicle screw instrumentation as a support for posterolateral spinal fusion: randomized clinical study with a 5-year follow-up. Spine (Phila Pa 1976) 2002;27:1269–77.
33. Christensen FB. Lumbar spinal fusion: outcome in relation to surgical methods, choice of implant and postoperative rehabilitation. Acta Orthop Scand Suppl 2004;75:2–43.
34. Christensen FB, Hansen ES, Eiskjaer SP, Hoy K, Helmig P, Neumann P, et al. Circumferential lumbar spinal fusion with brantigan cage versus posterolateral fusion with titanium cotrel-dubousset instrumentation: a prospective, randomized clinical study of 146 patients. Spine (Phila Pa 1976) 2002;27:2674–83.
35. Christensen FB, Laurberg I, Bunger CE. Importance of the back-cafe concept to rehabilitation after lumbar spinal fusion: a randomized clinical study with a 2-year follow-up. Spine (Phila Pa 1976) 2003;28:2561–9.
36. Videbaek TS, Christensen FB, Soegaard R, Hansen ES, Hoy K, Helmig P, et al. Circumferential fusion improves outcome in comparison with instrumented posterolateral fusion: long-term results of a randomized clinical trial. Spine (Phila Pa 1976) 2006;31:2875–80.
37. Nuhr MJ, Crevenna R, Quittan M, Auterith A, Wiesinger GF, Brockow T, et al. Cross-cultural adaption of the Manniche questionnaire for German-speaking low back pain patients. J Rehabil Med 2004;36:267–72.
38. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC Musculoskelet Disord 2006;7:82.
39. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry Low Back Pain Disability Questionnaire. Physiotherapy 1980;66:271–3.
40. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976) 2000;25:2940–52.
41. Roland M, Fairbank J. The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. Spine (Phila Pa 1976) 2000;25:3115–24.
42. Deyo RA, Battie M, Beurskens AJ, Bombardier C, Croft P, Koes B, et al. Outcome measures for low back pain research: a proposal for standardized use. Spine (Phila Pa 1976) 1998;23:2003–13.
43. Brox JI, Sorensen R, Friis A, Nygaard O, Indahl A, Keller A, et al. Randomized clinical trial of lumbar instrumented fusion and cognitive intervention and exercises in patients with chronic low back pain and disc degeneration. Spine (Phila Pa 1976) 2003;28:1913–21.
44. Carette S, Leclaire R, Marcoux S, Morin F, Blaise GA, St-Pierre A, et al. Epidural corticosteroid injections for sciatica due to herniated nucleus pulposus. N Engl J Med 1997;336:1634–40.
45. Fritzell P, Hagg O, Wessberg P, Nordwall A. 2001 Volvo Award Winner in Clinical Studies. Lumbar fusion versus nonsurgical treatment for chronic low back pain: a multicenter randomized controlled trial from the Swedish Lumbar Spine Study Group. Spine (Phila Pa 1976) 2001;26:2521–32.
46. Malmivaara A, Hakkinen U, Aro T, Heinrichs ML, Koskenniemi L, Kuosma E, et al. The treatment of acute low back pain: bed rest, exercises or ordinary activity. N Engl J Med 1995;332:351–5.
47. Weinstein JN, Tosteson TD, Lurie JD, Tosteson AN, Hanscom B, Skinner JS, et al. Surgical vs nonoperative treatment for lumbar disk

herniation: the Spine Patient Outcomes Research Trial (SPORT). A randomized trial. JAMA 2006;296:2441–50.

48. Costa LO, Maher CG, Latimer J. Self-report outcome measures for low back pain: searching for international cross-cultural adaptations. Spine (Phila Pa 1976) 2007;32:1028–37.

49. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 2002;82:8–24.

50. Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 2003;12:12–20.

51. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71–6.

52. Stratford PW, Binkley J, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. Phys Ther 1994;74:528–33.

53. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine (Phila Pa 1976) 2008;33:90–4.

54. Mayer TG, Barnes D, Kishino ND, Nichols G, Gatchel RJ, Mayer H, et al. Progressive isoinertial lifting evaluation. I. A standardized protocol and normative database. Spine (Phila Pa 1976) 1988;13:993–7.

55. Mayer TG, Barnes D, Nichols G, Kishino ND, Coval K, Piel B, et al. Progressive isoinertial lifting evaluation. II. A comparison with isokinetic lifting in a disabled chronic low-back pain industrial population. Spine (Phila Pa 1976) 1988;13:998–1002.

56. Lindell O, Eriksson L, Strender LE. The reliability of a 10-test package for patients with prolonged back and neck pain: could an examiner without formal medical education be used without loss of quality? A methodological study. BMC Musculoskelet Disord 2007;8:31.

57. Rainville J, Sobel J, Hartigan C, Monlux G, Bean J. Decreasing disability in chronic back pain through aggressive spine rehabilitation. J Rehabil Res Dev 1997;34:383–93.

58. Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, Van Der Heijden GJ, Knottnerus JA. Chronic low back pain: physical training, graded activity with problem solving training, or both? The one-year post-treatment results of a randomized controlled trial. Pain 2008;134:263–76.

59. Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, van der Heijden GJ, van Geel AC, et al. Active rehabilitation for chronic low back pain: cognitive-behavioral, physical, or both? First direct post-treatment results from a randomized controlled trial. BMC Musculoskelet Disord 2006; 7:5.

60. Weiner DK, Rudy TE, Glick RM, Boston JR, Lieber SJ, Morrow LA, et al. Efficacy of percutaneous electrical nerve stimulation for the treatment of chronic low back pain in older adults. J Am Geriatr Soc 2003;51:599–608.

61. Ljungquist T, Fransson B, Harms-Ringdahl K, Bjornham A, Nygren A. A physiotherapy test package for assessing back and neck dysfunction: discriminative ability for patients versus healthy control subjects. Physiother Res Int 1999;4:123–40.

62. Ljungquist T, Harms-Ringdahl K, Nygren A, Jensen I. Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. Physiother Res Int 1999;4:214–32.

63. Ljungquist T, Jensen IB, Nygren A, Harms-Ringdahl K. Physical performance tests for people with long-term spinal pain: aspects of construct validity. J Rehabil Med 2003;35:69–75.

64. Ljungquist T, Nygren A, Jensen I, Harms-Ringdahl K. Physical performance tests for people with spinal pain: sensitivity to change. Disabil Rehabil 2003;25:856–66.

65. Lygren H, Dragesund T, Joensen J, Ask T, Moe-Nilssen R. Test-retest reliability of the Progressive Isoinertial Lifting Evaluation (PILE). Spine (Phila Pa 1976) 2005;30:1070–4.

66. Andersson EI, Lin CC, Smeets RJ. Performance tests in people with chronic low back pain: responsiveness and minimal clinically important change. Spine (Phila Pa 1976) 2010;35:E1559–63.

67. Smeets RJ, Van Geel AC, Kester AD, Knottnerus JA. Physical capacity tasks in chronic low back pain: what is the contributing role of cardiovascular capacity, pain and psychological factors? Disabil Rehabil 2007;29:577–86.

68. Smeets RJ, Hijdra HJ, Kester AD, Hitters MW, Knottnerus JA. The usability of six physical performance tasks in a rehabilitation population with chronic low back pain. Clin Rehabil 2006;20:987–96.

69. Mayer T, Gatchel R, Mooney V. Safety of the dynamic Progressive Isoinertial Lifting Evaluation (PILE) test. Spine (Phila Pa 1976) 1990; 15:985–6.

70. Horneij E, Holmstrom E, Hemborg B, Isberg PE, Ekdahl C. Inter-rater reliability and between-days repeatability of eight physical performance tests. Adv Physiother 2002;4:146–60.

71. Curtis L, Mayer TG, Gatchel RJ. Physical progress and residual impairment quantification after functional restoration. Part III: isokinetic and isoinertial lifting capacity. Spine (Phila Pa 1976) 1994;19:401–5.

72. Soer R, Poels BJ, Geertzen JH, Reneman MF. A comparison of two lifting assessment approaches in patients with chronic low back pain. J Occup Rehabil 2006;16:639–46.

73. Lemstra M, Olszynski WP, Enright W. The sensitivity and specificity of functional capacity evaluations in determining maximal effort: a randomized trial. Spine (Phila Pa 1976) 2004;29:953–9.

74. Geisser ME, Haig AJ, Theisen ME. Activity avoidance and function in persons with chronic low back pain. J Occup Rehabil 2000;10:215–27.

75. Alschuler KN, Theisen-Goodvich ME, Haig AJ, Geisser ME. A comparison of the relationship between depression, perceived disability, and physical performance in persons with chronic pain. Eur J Pain 2008;12:757–64.

76. Schenk P, Klipstein A, Spillmann S, Stroyer J, Laubli T. The role of back muscle endurance, maximum force, balance and trunk rotation control regarding lifting capacity. Eur J Appl Physiol 2006;96:146–56.

77. Von Garnier K, Ewert T, Freumuth R, Limm H, Stucki G. Factors explaining improvement of isoinertial lifting capacity. J Occup Rehabil 2007;17:652–66.

78. Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, et al. The Quebec Back Pain Disability Scale: conceptualization and development. J Clin Epidemiol 1996;49:151–61.

79. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. Phys Ther 2001;81:776–88.

80. Cusi M, Saunders J, Hungerford B, Wisbey-Roth T, Lucas P, Wilson S. The use of prolotherapy in the sacroiliac joint. Br J Sports Med 2010;44:100–4.

81. Conway J, Tomkins CC, Haig AJ. Walking assessment in people with lumbar spinal stenosis: capacity, performance, and self-report measures. Spine J 2011. E-pub ahead of print.

82. Almeida DB, Prandini MN, Awamura Y, Vitola ML, Simiao MP, Milano JB, et al. Outcome following lumbar disc surgery: the role of fibrosis. Acta Neurochir (Wien) 2008;150:1167–76.

83. Louisia S, Anract P, Babinet A, Tomeno B, Revel M, Poiraudeau S. Long-term disability assessment after surgical treatment of low grade spondylolisthesis. J Spinal Disord 2001;14:411–6.

84. Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, et al. The Quebec Back Pain Disability Scale: measurement properties. Spine (Phila Pa 1976) 1995;20:341–52.

85. Verbunt JA, Sieben JM, Seelen HA, Vlaeyen JW, Bousema EJ, van der Heijden GJ, et al. Decline in physical activity, disability and pain-related fear in sub-acute low back pain. Eur J Pain 2005;9:417–25.

86. Sanchez K, Papelard A, Nguyen C, Jousse M, Rannou F, Revel M, et al. Patient-preference disability assessment for disabling chronic low back pain: a cross-sectional survey. Spine (Phila Pa 1976) 2009;34: 1052–9.

87. Reneman MF, Jorritsma W, Schellekens JM, Goeken LN. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. J Occup Rehabil 2002;12:119–29.

88. Van den Hout JH, Vlaeyen JW, Heuts PH, Zijlema JH, Wijnen JA. Functional disability in nonspecific low back pain: the role of pain-related fear and problem-solving skills. Int J Behav Med 2001;8:134–48.

89. Wilhelm F, Fayolle-Minon I, Phaner V, Le-Quang B, Rimaud D, Bethoux F, et al. Sensitivity to change of the Quebec Back Pain Disability Scale and the Dallas Pain Questionnaire. Ann Phys Rehabil Med 2010;53:15–23.

90. Schoppink Le, van Tulder MW, Koes BW, Beurskens SA, de Bie RA. Reliability and validity of the Dutch adaptation of the Quebec Back Pain Disability Scale. Phys Ther 1996;76:268–75.

91. Yvanes-Thomas M, Calmels P, Bethoux F, Richard A, Nayme P, Payre D, et al. Validity of the French-language version of the Quebec Back Pain Disability Scale in low back pain patients in France. Joint Bone Spine 2002;69:397–405.

92. Mousavi SJ, Parnianpour M, Mehdian H, Montazeri A, Mobini B. The Oswestry Disability Index, the Roland-Morris Disability Questionnaire, and the Quebec Back Pain Disability Scale: translation and validation studies of the Iranian versions. Spine (Phila Pa 1976) 2006; 31:E454–9.

93. Rodrigues MF, Michel-Crosato E, Cardoso JR, Traebert J. Psychometric properties and cross-cultural adaptation of the Brazilian Quebec Back Pain Disability Scale Questionnaire. Spine (Phila Pa 1976) 2009;34: E459–64.

94. Bicer A, Yazici A, Camdeviren H, Milcan A, Erdogan C. Assessment of pain and disability in patients with chronic low back pain: reliability and construct validity of the Turkish version of the Quebec Back Pain Disability Scale and Pain Disability Index. J Back Musculoskelet Rehabil 2005;18:37–44.

95. Melikoglu MA, Kocabas H, Sezer I, Bilgilisoy M, Tuncer T. Validation of the Turkish version of the Quebec Back Pain Disability Scale for patients with low back pain. Spine (Phila Pa 1976) 2009;34:E219–24.

96. Bendeddouche I, Rostom S, Bahiri R, Boudali A, Mawani N, Mengat M, et al. Traduction, adaptation transculturelle et validation de la version marociane de la Quebec Back Pain Disability Scale [abstract]. Rev Rhum 2010;77 Suppl:A179.
97. Demoulin C, Ostelo R, Knottnerus JA, Smeets RJ. Quebec Back Pain Disability Scale was responsive and showed reasonable interpretability after a multidisciplinary treatment. J Clin Epidemiol 2010;63: 1249–55.
98. Hicks GE, Manal TJ. Psychometric properties of commonly used low back disability questionnaires: are they useful for older adults with low back pain? Pain Med 2009;10:85–94.
99. Rocchi MB, Sisti D, Benedetti P, Valentini M, Bellagamba S, Federici A. Critical comparison of nine different self-administered questionnaires for the evaluation of disability caused by low back pain. Eura Medicophys 2005;41:275–81.
100. Van der Roer N, Ostelo RW, Bekkering GE, van Tulder MW, de Vet HC. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine (Phila Pa 1976) 2006;31:578–82.
101. Roland M, Morris R. A study of the natural history of back pain part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976) 1983;8:141–4.
102. Grotle M, Brox JI, Vollestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. Spine (Phila Pa 1976) 2004;29:E492–501.
103. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine (Phila Pa 1976) 1995;20:1899–909.
104. Stratford PW, Binkley JM. Measurement properties of the RM-18: a modified version of the Roland-Morris Disability Scale. Spine (Phila Pa 1976) 1997;22:2416–21.
105. Underwood MR, Barnett AG, Vickers MR. Evaluation of two time-specific back pain outcome measures. Spine (Phila Pa 1976) 1999;24: 1104–12.
106. Kopec JA, Esdaile JM. Functional disability scales for back pain. Spine (Phila Pa 1976) 1995;20:1943–9.
107. Williams RM, Myers AM. A new approach to measuring recovery in injured workers with acute low back pain: Resumption of Activities of Daily Living Scale. Phys Ther 1998;78:613–23.
108. Garratt AM. Rasch analysis of the Roland Disability Questionnaire. Spine (Phila Pa 1976) 2003;28:79–84.
109. Turner JA, Fulton-Kehoe D, Franklin G, Wickizer TM, Wu R. Comparison of the Roland-Morris Disability Questionnaire and generic health status measures: a population-based study of workers' compensation back injury claimants. Spine (Phila Pa 1976) 2003;28:1061–7.
110. Artus M, van der Windt DA, Jordan KP, Hay EM. Low back pain symptoms show a similar pattern of improvement following a wide range of primary care treatments: a systematic review of randomized clinical trials. Rheumatology (Oxford) 2010;49:2346–56.
111. Wilkens P, Scheel IB, Grundnes O, Hellum C, Storheim K. Effect of glucosamine on pain-related disability in patients with chronic low back pain and degenerative lumbar osteoarthritis: a randomized controlled trial. JAMA 2010;304:45–52.
112. Lamb SE, Hansen Z, Lall R, Castelnuovo E, Withers EJ, Nichols V, et al. Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis. Lancet 2010;375:916–23.
113. Hay EM, Mullis R, Lewis M, Vohora K, Main CJ, Watson P, et al. Comparison of physical treatments versus a brief pain-management programme for back pain in primary care: a randomised clinical trial in physiotherapy practice. Lancet 2005;365:2024–30.
114. Mannion AF, Muntener M, Taimela S, Dvorak J. A randomized clinical trial of three active therapies for chronic low back pain. Spine (Phila Pa 1976) 1999;24:2435–48.
115. Bishop FL, Lewis G, Harris S, McKay N, Prentice P, Thiel H, et al. A within-subjects trial to test the equivalence of online and paper outcome measures: the Roland Morris Disability Questionnaire. BMC Musculoskelet Disord 2010;11:113.
116. Davidson M. Rasch analysis of 24-, 18- and 11-item versions of the Roland-Morris Disability Questionnaire. Qual Life Res 2009;18:473–81.
117. Kucukdeveci AA, Tennant A, Elhan AH, Niyazoglu H. Validation of the Turkish version of the Roland-Morris Disability Questionnaire for use in low back pain. Spine (Phila Pa 1976) 2001;26:2738–43.
118. Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable change for the Roland-Morris Questionnaire. Phys Ther 1996;76:359–68.
119. Beurskens AJ, de Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain: assessment of the quality of four disease-specific questionnaires. Spine (Phila Pa 1976) 1995;20:1017–28.
120. Albert HB, Jensen AM, Dahl D, Rasmussen MN. Criteria validation of the Roland Morris Questionnaire: a Danish translation of the international scale for the assessment of functional level in patients with low back pain and sciatica. Ugeskr Laeger 2003;165:1875–80. In Danish.
121. Gommans I, Koes B, van Tulder M. Validity and responsiveness of the Dutch version of the Roland Disability Questionnaire: a functional status questionnaire for patients with low back pain. Ned Tijds Fysioth 1997;107:28–33.
122. Coste J, Le Parc JM, Berge E, Delecoeuillerie G, Paolaggi JB. French validation of a disability rating scale for the evaluation of low back pain (EIFEL questionnaire). Rev Rhum Ed Fr 1993;60:335–41. In French.
123. Wiesinger GF, Nuhr M, Quittan M, Ebenbichler G, Wolfl G, Fialka-Moser V. Cross-cultural adaptation of the Roland-Morris Questionnaire for German-speaking patients with low back pain. Spine (Phila Pa 1976) 1999;24:1099–103.
124. Boscainos PJ, Sapkas G, Stilianessi E, Prouskas K, Papadakis SA. Greek versions of the Oswestry and Roland-Morris Disability Questionnaires. Clin Orthop Relat Res 2003;411:40–53.
125. Padua R, Padua L, Ceccarelli E, Romanini E, Zanoli G, Bondi R, et al. Italian version of the Roland Disability Questionnaire, specific for low back pain: cross-cultural adaptation and validation. Eur Spine J 2002; 11:126–9.
126. Suzukamo Y, Fukuhara S, Kikuchi S, Konno S, Roland M, Iwamoto Y, et al. Validation of the Japanese version of the Roland-Morris Disability Questionnaire. J Orthop Sci 2003;8:543–8.
127. Grotle M, Brox JI, Vollestad NK. Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. J Rehabil Med 2003;35:241–7.
128. Costa LO, Maher CG, Latimer J, Ferreira PH, Pozzi GC, Ribeiro RN. Psychometric characteristics of the Brazilian-Portuguese versions of the Functional Rating Index and the Roland Morris Disability Questionnaire. Spine (Phila Pa 1976) 2007;32:1902–7.
129. Nusbaum L, Natour J, Ferraz MB, Goldenberg J. Translation, adaptation and validation of the Roland-Morris Questionnaire: Brazil Roland-Morris. Braz J Med Biol Res 2001;34:203–10.
130. Maaroufi H, Benbouazza K, Faik A, Bahiri R, Lazrak N, Abouqal R, et al. Translation, adaptation, and validation of the Moroccan version of the Roland Morris Disability Questionnaire. Spine (Phila Pa 1976) 2007;32:1461–5.
131. Kovacs FM, Llobera J, Gil Del Real MT, Abraira V, Gestoso M, Fernandez C, et al. Validation of the Spanish version of the Roland-Morris Questionnaire. Spine (Phila Pa 1976) 2002;27:538–42.
132. Scharovsky A, Pueyrredon M, Craig D, Rivas ME, Converso G, Pueyrredon JH, et al. Cross-cultural adaptation and validation of the Argentinean version of the Roland-Morris Disability Questionnaire. Spine (Phila Pa 1976) 2008;33:1391–5.
133. Johansson E, Lindberg P. Subacute and chronic low back pain: reliability and validity of a Swedish version of the Roland and Morris Disability Questionnaire. Scand J Rehabil Med 1998;30:139–43.
134. Bejia I, Abid N, Ben Salem K, Letaief M, Younes M, Touzi M, et al. Low back pain in a cohort of 622 Tunisian schoolchildren and adolescents: an epidemiological study. Eur Spine J 2005;14:331–6.
135. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: development and final revision of a health status measure. Med Care 1981;19:787–805.
136. Costa LO, Maher CG, Latimer J, Ferreira PH, Ferreira ML, Pozzi GC, et al. Clinimetric testing of three self-report outcome measures for low back pain patients in Brazil: which one is the best? Spine (Phila Pa 1976) 2008;33:2459–63.
137. Frost H, Lamb SE, Stewart-Brown S. Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index V2.1 and Roland and Morris Disability Questionnaire for patients with subacute and chronic low back pain. Spine (Phila Pa 1976) 2008;33:2450–8.
138. Jarvikoski A, Mellin G, Estlander A. Outcome of two multimodal back treatment programmes with and without intensive physical training. J Spinal Disord 1995;6:93–8.
139. Hsieh CY, Phillips RB, Adams AH, Pope MH. Functional outcomes of low back pain: comparison of four treatment groups in a randomized controlled trial. J Manipulative Physiol Ther 1992;15:4–9.
140. Jensen MP, Strom SE, Turner JA, Romano JM. Validity of the Sickness Impact Profile Roland Scale as a measure of dysfunction in chronic pain patients. Pain 1992;50:157–62.
141. Deyo RA. Comparative validity of the sickness impact profile and shorter scales for functional assessment in low-back pain. Spine (Phila Pa 1976) 1986;11:951–4.
142. Brouwer S, Kuijer W, Dijkstra PU, Goeken LN, Groothoff JW, Geertzen JH. Reliability and stability of the Roland Morris Disability Questionnaire: intra class correlation and limits of agreement. Disabil Rehabil 2004;26:162–5.
143. Demoulin C, Ostelo R, Knottnerus JA, Smeets RJ. what factors influence the measurement properties of the Roland-Morris Disability Questionnaire? Eur J Pain 2010;14:200–6.
144. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the

Roland-Morris Back Pain Questionnaire: part 2. Phys Ther 1998;78: 1197–207.

145. Kovacs FM, Abraira V, Royuela A, Corcoll J, Alegre L, Cano A, et al. Minimal clinically important change for pain intensity and disability in patients with nonspecific low back pain. Spine (Phila Pa 1976) 2007;32:2915–20.

146. Pengel LH, Refshauge KM, Maher CG. Responsiveness of pain, disability, and physical impairment outcomes in patients with low back pain. Spine (Phila Pa 1976) 2004;29:879–83.

147. Leclaire R, Blier F, Fortin L, Proulx R. A cross-sectional study comparing the Oswestry and Roland-Morris Functional Disability Scales in two populations of patients with low back pain of different levels of severity. Spine (Phila Pa 1976) 1997;22:68–71.

148. Stratford PW, Binkley JM. A comparison study of The Back Pain Functional Scale and Roland Morris Questionnaire: North American Orthopaedic Rehabilitation Research Network. J Rheumatol 2000;27: 1928–36.

149. Garratt AM, Klaber Moffett J, Farrin AJ. Responsiveness of generic and specific measures of health outcome in low back pain. Spine (Phila Pa 1976) 2001;26:71–7.

150. Kovacs FM, Abraira V, Zamora J, Teresa Gil del Real M, Llobera J, Fernandez C, et al. correlation between pain, disability, and quality of life in patients with common low back pain. Spine (Phila Pa 1976) 2004;29:206–10.

151. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39:897–906.

152. Ekedahl KH, Jonsson B, Frobell RB. Validity of the fingertip-to-floor test and straight leg raising test in patients with acute and subacute low back pain: a comparison by sex and radicular pain. Arch Phys Med Rehabil 2010;91:1243–7.

153. Kuijer W, Brouwer S, Dijkstra PU, Jorritsma W, Groothoff JW, Geertzen JH. Responsiveness of the Roland-Morris Disability Questionnaire: consequences of using different external criteria. Clin Rehabil 2005;19:488–95.

154. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther 1998;78:1186–96.

155. Hare-Mortensen L, Lauridsen H, Grunnet-Nilsson N. The relative responsiveness of 3 different types of clinical outcome measures on chiropractic patients with low back pain. J Manipulative Physiol Ther 2006;29:95–9.

Summary Table for Measures of Function in Low Back Pain/Disorders*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
PILE	Capacity to tolerate strenuous lifting throughout a day and to evaluate lifting capacity	Observer-led task	5–15 min, temporary increase of pain	5–15 min, increase of lifting weight, register time, HR, and quality of lifting	Weight lifted adjusted for sex/weight or number of completed lifting cycles	Good ICCs; however, LOA large (48% of baseline score)	Good and proof for construct validity	Poor to moderate	Safe, inexpensive, easy to administer psychophysical lifting end point, and unconstrained lifting reflects “real-world” lifting	N/A in patients taking HR-limiting medication. Unable to discriminate the “weak link” of the biomechanical lifting chain 10% of patients are not able to perform task
ODI	Measuring pain-related disability in people with acute, subacute, or chronic low back pain	Self-completed questionnaire by patient on paper and/or phone	<5 min	<1 min	Total score ranges from 0 (no disability) to 100 (maximum disability)	Good ICCs	Adequate content and construct validity; however, lacks generic activities such as work, leisure, recreational, or sporting activities	Cutoff point for minimum important change is 10 points or a 30% score improvement	Simple to use and score, and has minimal respondent and administrative burden	Face-to-face or computer administration would be the preferred method over telephone interview
LBPRS	Measuring 3 clinical illness components of low back pain: pain (back and leg), disability, and physical impairment	Self-completed questionnaire by patient on paper or by interview	~15 min	~15 min	Score ranges: 0–60 for pain, 0–40 points for disability, 0–40 points for impairments Recommended not to use the total sum score Higher scores are indicative of more problems	High interrater reliability (97.7%)	Correlates highly with RDQ	MCID for the disability scale is 17 and for pain scale is 10 points	Simple and contains a well-balanced distribution of items across the ICF components pain, activity limitation, and physical impairment	Responsiveness is lower compared to RDQ and ODI Lacks information on MDC and SEM
RDQ	Measuring daily physical activities and functions that may be affected by low back pain	Self-completed questionnaire by patient on paper and electronic version	<5 min	<1 min	Scores range from 0 (no disability) to 24 (maximal disability)	Internal consistency and ICC are good MDC and SEM are known, but are influenced by several factors (time intervals, methods used, etc.)	Acceptable; contains a small number of items that are not related to functional limitations Correlates well with other disability measures	MCID ranges from 2–5 points. A 30% change from baseline was proposed as a clinically meaningful improvement (normally equivalent to an absolute change of 5 points)	Short, simple to complete, and readily understood by patients and clinicians. Psychometric properties are acceptable to good and the RDQ is available in many language versions. It can be used in acute, subacute, and chronic low back pain patients	Less suitable for patients with low levels of disability. Can be improved through the removal of items with poor fit statistics and the addition of items toward the extremes of the scale hierarchy
QBPDS	Measuring elementary daily activities that patients with back pain might perceive difficult to perform. Items can be classified into 6 domains of activity affected by back pain	Self-completed questionnaire by patient on paper, mail, and/or phone	<5 min	<1 min	Score ranges from 0 (no disability) to 100 (maximal disability)	Internal consistency and ICC are good	Good, contains various domains of activity that were selected by patients and health care providers; correlates well with other disability measures	MCID ranges from 8.5–32.9 mainly due to the heterogeneity of the study populations. A 30% change from baseline was proposed as a clinically meaningful improvement	Short, easy to use, and acceptable. Measures functional disability in daily life that is essential in patients with low back pain	Due to changes regarding the scale's format and the wording of some of the items, one cannot be sure that all clinimetric properties reported in studies are identical for the newly proposed version

* PILE = Progressive Isoinertial Lifting Evaluation; HR = heart rate; ICC = intraclass correlation coefficient; LOA = limits of agreement; N/A = not applicable; ODI = Oswestry Disability Index; LBPRS = Low Back Pain Rating Scale; RDQ = Roland-Morris Disability Questionnaire; MCID = minimum clinically important difference; ICF = International Classification of Functioning, Disability and Health; MDC = minimum detectable change; QBPDS = Quebec Back Pain Disability Scale.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S337–S349
DOI 10.1002/acr.20633

Measures of Work Disability and Productivity

Rheumatoid Arthritis Specific Work Productivity Survey (WPS-RA), Workplace Activity Limitations Scale (WALS), Work Instability Scale for Rheumatoid Arthritis (RA-WIS), Work Limitations Questionnaire (WLQ), and Work Productivity and Activity Impairment Questionnaire (WPAI)

KENNETH TANG,[1] DORCAS E. BEATON,[1] ANNELIES BOONEN,[2] MONIQUE A. M. GIGNAC,[3] AND CLAIRE BOMBARDIER[4]

INTRODUCTION

The impact of arthritis on work is an area of increasing research interest and a growing number of outcome measures to quantify such impact have become available in recent years. Recent reviews from an Outcome Measures in Rheumatology initiative have broadly identified 24 instruments in this area (1,2), though only 11 had been used in arthritis to date. Previous studies have shown that these measures only moderately correlate with each other (3–10), therefore it is important to recognize that available instruments offer distinct perspectives on health-related work impacts. For example, some measures are aimed at examining degree of difficulties with specific workplace activities, while others are designed to quantify the extent of absenteeism (e.g., number of days off work) and/or presenteeism (e.g., being at work but working at reduced productivity, also referred to as "at-work productivity loss" or "at-work disability"). Also available are instruments focused on assessing related concepts such as work "performance," "efficiency," "instability," or degree of "interference" at work. For the purpose of this review, we have adopted a broader approach and consider a diverse range of available measures that offer varying perspectives and approaches to quantifying the impact of health problems on work (Table 1).

Beyond their diverse conceptual foci, existing instruments also differ in terms of their scope of measurement (e.g., impact on employment work versus nonpaid work and/or leisure activities), disease attribution (e.g., disease specific versus generic), length (e.g., number of sections and items), and recall period. Some are designed as "modular" instruments that assess work impacts using a series of global rating scales and discrete items (often organized into multiple sections) that are generally not intended to be summative. Others are classic "psychometric" measures consisting of summative items that inform disease impact on different specific aspects of work (contributes to an overall construct). Specific impacts examined at the item level may include problems meeting the physical demands of work, challenges associated with time management, difficulties maintaining interpersonal relationships at work, cognitive concerns (e.g., worries about continued employability), and/or issues related to symptom control/exacerbation and fatigue. The diversity of available measures is a strength in this growing field and has accommodated the growing interest to apply these tools for a broad range of purposes in arthritis. For example, these measures have been used to examine the epidemiology (e.g., population trends, determinants) of work disability/participation in the arthritis population, to evaluate the effectiveness of clinical or workplace interventions (e.g., clinical trials), and also, to estimate the economic costs of health-related work productivity loss at the societal level.

Five specific measures were selected for a detailed review in this article. These were chosen on the basis of 2

Mr. Tang is recipient of a Canadian Institutes of Health Research PhD Fellowship, Canadian Arthritis Network Graduate Award, and Syme Fellowship from the Institute for Work & Health. Dr. Beaton is recipient of a Canadian Institutes of Health Research New Investigators award. Dr. Bombardier is recipient of a Canada Research Chair in Knowledge Transfer for Musculoskeletal Care.

[1]Kenneth Tang, MSc(PT), MSc, Dorcas E. Beaton, PhD: St. Michael's Hospital, Institute for Work & Health, and University of Toronto, Toronto, Ontario, Canada; [2]Annelies Boonen, MD, PhD: Maastricht University Medical Center and Caphri Research Institute, Maastricht, The Netherlands; [3]Monique A. M. Gignac, PhD: Toronto Western Research Institute, University Health Network, and University of Toronto, Toronto, Ontario, Canada; [4]Claire Bombardier, MD: University Health Network, Institute for Work & Health, and University of Toronto, Toronto, Ontario, Canada.

Dr. Boonen has received consultancy fees, speaking fees, and/or honoraria (less than $10,000 each) from Abbott and Pfizer.

Address correspondence to Kenneth Tang, MSc(PT), MSc, Mobility Program Clinical Research Unit, Li Ka Shing Institute of St. Michael's Hospital, 30 Bond Street, Toronto, Ontario, Canada. E-mail: ken.tang@utoronto.ca.

Submitted for publication January 31, 2011; accepted in revised form May 10, 2011.

Table 1. Summary of measures of role functioning and productivity at work, sorted by year of original publication

Author, year (ref.)	Measure	Measure type*	No. items/ sections†	Scope‡
Osterhaus et al, 1992 (18)	Osterhaus Technique (OST)	M	4	A, P
Reilly et al, 1993 (15)	Work Productivity and Activity Impairment-General Health (WPAI)§	M	6	A, P
Van Roijen et al, 1996 (19)	Health and Labor Questionnaire (HLQ)	M	14	A, P
Endicott et al, 1997 (20)	Endicott Work Productivity Scale (EWPS)	Ps	25	P
Kopec et al, 1998 (21)	Occupational Role Questionnaire (ORQ)	Ps	8	P
Brouwer et al, 1999 (9)	Quantity and Quality Method (QQ) from the Productivity and Disease Questionnaire	M	2	P
Amick et al, 2000 (22)	Work Role Functioning-26 (WRF)	Ps	26	P
Lerner et al, 2001 (14)	Work Limitations Questionnaire-25 (WLQ-25)§	Ps	25	P
Altshuler et al, 2002 (23)	Life Functioning Questionnaire (LFQ)	M	16	A, P
Koopman et al, 2002 (24)	Stanford Presenteeism Scale-6 (SPS-6)	Ps	6	P
Kumar et al, 2003 (25)	Health-Related Productivity Questionnaire Diary (HRPQ-D)	M	9	A, P
Kessler et al, 2003 (26)	World Health Organization Health and Work Performance Questionnaire (HPQ)§	M	13	A, P
Gilworth et al, 2003 (13)	Work Instability Scale for Rheumatoid Arthritis (RA-WIS)§	Ps	23	P
Goetzel et al, 2003 (27)	Work Productivity Short Inventory (Wellness Inventory by Pfizer) (WPSI)	M	4	A, P
Shikiar et al, 2004 (28)	Health and Work Questionnaire (HWQ)	M	5	P
Turpin et al, 2004 (6)	Stanford Presenteeism Scale-13 (SPS-13)	Ps	13	A, P
Gignac et al, 2004 (12)	Work Activity Limitations Scale (WALS)§	Ps	12	P
Stewart et al, 2004 (29)	Work and Health Interview-The American Productivity Audit (WHI)	Ps	7	A, P
Burton et al, 2004 (30,31)	Work Limitations Questionnaire-8 (WLQ-8)	Ps	8	P
Beaton et al, 2005 (32)	Work Limitations Questionnaire-16 (WLQ-16)	Ps	16	P
Munir et al, 2005 (33)	Work Limitations Questionnaire, 6 items (WLQm)	Ps	6	P
Feuerstein et al, 2005 (34)	Workstyle Scale-Long Version (WSL)	Ps	91	P
Feuerstein et al, 2006 (35)	Workstyle Scale-Short Version (WSS)	Ps	32	P
Van Roijen, 2007 (36)	Short Form-Health Labor Questionnaire (SF-HLQ)	M	11	A, P
Osterhaus et al, 2009 (11)	Rheumatoid Arthritis Specific Work Productivity Survey (WPS-RA)§	M	9	A, P

* Modular measure (M) consists of series of global rating scales (e.g., visual analog scales or numeric rating scales) and discrete items generally not intended to be summative; contents may be organized into multiple sections within a questionnaire. Psychometric measure (Ps) consists of summative items that contributes to overall concept (can be multidimensional, i.e., consisting of subscales).

† For modular measures organized by sections, the number of sections (not items) are indicated; count excludes sociodemographic items/sections unless directly pertaining to employment work (e.g., current work status, employment income).

‡ Absenteeism (A) examines work status and/or extent of time/frequency of being off work (e.g., sick leave). Presenteeism (P) examines on-the-job impact (e.g., productivity loss associated with reduced work efficiency, or degree of workplace activity limitations).

§ Identified as a candidate Outcome Measures in Rheumatology work productivity outcome measure (2).

main criteria: the availability of measurement evidence specific to the arthritis population (primarily osteoarthritis or inflammatory arthritis), and evidence of previous research application in arthritis populations beyond psychometric testing. By these selection criteria, both modular and psychometric measures were represented. Measures chosen included the Rheumatoid Arthritis Specific Work Productivity Survey (11), Workplace Activity Limitations Scale (12), Work Instability Scale for Rheumatoid Arthritis (13), Work Limitations Questionnaire (14), and Work Productivity and Activity Impairment Questionnaire (15). It should also be recognized that disease-specific variants of 2 of the 5 selected measures are available for ankylosing spondylitis (i.e., the Work Productivity and Activity Impairment Questionnaire for Ankylosing Spondylitis [16], and the Ankylosing Spondylitis Work Instability Scale [17]); however, these measures will not be presented in this review. Also, we have emphasized psychometric evidence primarily in arthritis populations, notwithstanding that evidence in nonarthritis soft tissue musculoskeletal conditions (e.g., low back pain, upper-extremity disorders) and other nonmusculoskeletal disorders are also available for a number of these measures.

RHEUMATOID ARTHRITIS SPECIFIC WORK PRODUCTIVITY SURVEY (WPS-RA)

Description

Purpose. Measures the impact of rheumatoid arthritis (RA) on the productivity of employment work, household work, and daily activities. The original version of WPS-RA was published in 2009, primarily intended for use in clinical trials (11).

Content. Assesses the number of days of work absence (absenteeism), number of days with reduced work productivity, and degree of interference on work productivity due to RA (presenteeism).

Number of items. 9 questions (Q), organized into 3 sections. Section I: employment status, type of work (Q1); section II: employment work (days of work missed [Q2], days with productivity reduced by at least half [Q3], degree of interference on work productivity [Q4]); and section III: household work and activities (days of nonpaid work missed [Q5], days with productivity reduced by at least half [Q6], days missed family, social, or leisure activities [Q7], days with outside help [Q8], degree of interference on [nonpaid] work productivity [Q9]).

Response options/scale. Unique for each individual question. Q1: employed outside of home (yes/no), type of work (8 options); Q2, Q3, and Q5–Q8: number of days (count data); Q4 and Q9: global rating scale, 0–10 (0 = no interference, 10 = complete interference).

Recall period for items. 1 month.

Endorsements. None. (The WPS-RA is 1 of 6 measures identified as a candidate Outcome Measures in Rheumatology work productivity outcome.)

Examples of use. The WPS-RA has been applied as a secondary study end point to evaluate effects of certolizumab pegol in RA (37,11), and also as an outcome to evaluate the effects of certolizumab pegol with methotrexate in RA (38).

Practical Application

How to obtain. Request for permission to use the WPS-RA should be made to Global Health Outcomes Research, UCB Pharma. A copy of the WPS-RA questionnaire can be obtained at URL: http://www.biomedcentral.com/content/supplementary/ar2702-S1.doc.

Method of administration. Currently intended for interviewer administration, but there are future plans to develop a self-administered version.

Scoring. Each of 9 questions is scored individually (i.e., 8 separate outcome scores [Q1 not considered an outcome]), item scores are not intended to be combined.

Score interpretation. Values for Q2, Q3, and Q5–Q8 reflect number of days impacted; for Q4 and Q9, higher scores indicate greater interference of RA on work. Cut points and normative values are not yet established.

Respondent burden. Low.

Administrative burden. Low, assuming interviewer administration.

Translations/adaptations. English.

Psychometric Information

Method of development. The WPS-RA was developed through a literature review on work productivity associated with RA or other chronic health conditions (e.g., migraine, depression) (11). Patients/workers were not directly involved in scale development, but item selection emphasized patient-centered considerations (e.g., relevance, burden).

Acceptability. In a randomized trial of 220 patients with RA, Osterhaus et al (11) reported relatively low missing frequencies of 0.5% (Q6), 1.4% (Q8), 1.8% (Q9), and 0% for all other scale items. Q2, Q3, and Q5–Q8 provide count data and may have propensity for floor effect. Readability of WPS-RA items is high.

Reliability. Test–retest or interrater evidence not yet available, internal consistency testing is not applicable.

Validity. Modest evidence to date. Relationship of the WPS-RA with other work-specific measures has yet to be assessed, but evidence of known group differences against general health indicators is available. In Osterhaus et al (11), WPS-RA item scores for Q2–Q4 (employment work) were shown to differ ($P < 0.05$) between known groups (≥ third quartile versus ≤ first quartile) based on scores from the Health Assessment Questionnaire disability index (HAQ DI), Short Form-36 physical component summary score (SF-36 PCS), and the SF-36 mental component summary score (MCS). Item scores for Q5–Q9 (household work and daily activities) were also shown to differ (mostly $P < 0.05$) between known groups (≥ third quartile versus ≤ first quartile) based on scores from the HAQ DI, SF-36 PCS, and SF-36 MCS.

Ability to detect change. There are no known reports to date on the sensitivity of the WPS-RA to known changes in work disability or productivity, but effect sizes have been assessed in persons with RA in a 24-week certolizumab pegol trial (11). For clinical responders based on the American College of Rheumatology (ACR20) criteria for 20% improvement (i.e., individuals showing a large change), standardized response mean (SRM) >0.8 (large effect size) was observed for Q4, Q6, and Q9, SRM 0.5–0.8 (moderate effect size) was shown for Q3 and Q5, and SRM <0.5 (small effect size) was shown for Q2, Q7, and Q8. For ACR20 nonresponders (i.e., individuals showing small change, no change, or deterioration), SRM <0.5 was evident for all WPS-RA items (Q2–Q9). In the same trial, clinical responders based on improvements in the HAQ DI (minimum clinically important difference [MCID] 0.22, i.e., a large change in disability) had an SRM >0.8 for Q4 and Q6, SRM >0.5 for Q3 and Q5, and SRM <0.5 for Q2, Q7, and Q8. Among HAQ DI nonresponders (i.e., small change, no change, or deterioration), SRM <0.5 was observed for all items (Q2–Q9). Some care is needed when interpreting SRMs reported for nonresponders as these are derived from assessing a pool of individuals who have undergone varying degrees of change in the trial.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Provides broad coverage of impacts on several work domains; questionnaire considers work impacts from the perspectives of both absenteeism and presenteeism (at-work productivity) as well as impacts on both employment work and nonpaid activities (e.g., household work, social and leisure activities).

Caveats and cautions. Patients/workers were not directly involved in conceptualization or scale development. Count data (number of days) are gathered for 6 of the 9 questionnaire items (proper statistical treatment required). Evidence to support application beyond RA is not yet available. There is a large number of "outcomes" (i.e., 8 separate scores derived, not intended to be summated/

combined). Q2–Q4 pertains specifically to employment work, therefore will not be relevant for persons who are temporarily unemployed at the time of data collection (i.e., lead to missing values). Number of "days with productivity reduced by at least half" (Q3 and Q6) could be challenging to appraise and may overlook days where there is also considerable impact but productivity is reduced by less than half. Overall, there is limited psychometric testing against work-specific indicators (to support validity and responsiveness) to date, but it should be considered that the WPS-RA is a relatively new measure with much promise.

Clinical usability. Good potential, although more studies to establish clinical parameters (e.g., MCID, patient acceptable symptomatic state) are needed. The WPS-RA could be potentially useful for providing a more complete view of disease impact (i.e., on both paid and unpaid work) to facilitate decision making on clinical management and issues around job modifications and work/life balance. It is also a highly feasible measure given low respondent and administrative burden.

Research usability. Research usability is promising. There is emerging evidence of its application and psychometric performance from clinical trials in RA. Low administrative and respondent burden.

WORKPLACE ACTIVITY LIMITATIONS SCALE (WALS)

Description

Purpose. Measures limitations experienced while performing workplace activities, including difficulties associated with upper-extremity functioning, lower-extremity functioning, concentration at work, and the pace and scheduling of work. The WALS is intended for arthritis populations. Original publication in 2004 (12).

Content. Assesses difficulties with mobility, prolonged sitting and standing, lifting, working with hands, crouching, bending or kneeling, reaching, scheduling, work hours, pace of work, concentration, and meeting current job demands. Respondents are asked to respond to questions assuming that assistance from others or gadgets and equipment are not available.

Number of items. 11-item and 12-item versions of the WALS are available. An item asking about "difficulties concentrating on work" is not included in the 11-item version. This item was added to the 12-item version based on patient feedback.

Response options/scale. Four-point Likert scaling: from no difficulty (score = 0) to not able to do (score = 3). Not applicable to my job and difficulty unrelated to arthritis response options are also available (both scored 0).

Recall period for items. "In general" or "typically."

Endorsements. None. (The WALS is 1 of 6 measures identified as a candidate Outcome Measures in Rheumatology work productivity outcome.)

Examples of use. Has been used in samples with inflammatory arthritis (IA; e.g., rheumatoid arthritis [RA], psoriatic arthritis), osteoarthritis (OA), and lupus. Examined as a factor associated with arthritis-related work changes, work transitions and job accommodations (12,39–41), behavioral coping efforts (42), arthritis–work spillover (43), chronic job stress and strain (44), and disclosure of chronic disease in the workplace (45).

Practical Application

How to obtain. Information on the WALS can be obtained free of charge, from Monique Gignac, PhD, Toronto Western Research Institute at the University Health Network, Toronto, Ontario, Canada. E-mail: gignac@uhnres.utoronto.ca.

Method of administration. Self- or interviewer administration.

Scoring. Can be expressed as mean of all scale items (range 0–3) or as a summed total score (summed score range 0–33 for 11-item version or 0–36 for 12-item version); mean values can be imputed for up to 2 missing items.

Score interpretation. Higher scores indicate greater workplace activity limitations. Preliminary data on cutoff values have been examined (41).

Respondent burden. Low.

Administrative burden. Minimal.

Translations/adaptations. English.

Psychometric Information

Method of development. Items were based on a review of the literature and were modeled after the Health Assessment Questionnaire (HAQ) and modified to be specific to workplace tasks and activities. Patients/workers were not directly involved in the development of the scale.

Acceptability. In a sample of 250 workers with either RA or OA, Beaton et al (3) reported WALS summed scores to be available in 234 patients (6% completed <10 of the 11 scale items) with 0% at floor score (WALS = 0) and 3.4% at ceiling score (WALS = 33). Dhanhani (40) reported <1% missing values for employed and not employed groups with lupus. Readability of WALS items is high.

Reliability. For the 11-item WALS, Gignac et al (12,43) and Gignac (42) reported a Cronbach's alpha range from 0.78–0.81 (n = 349–491 with OA or IA) over 4 time points, each 18 months apart. In a sample of 250 workers with either RA or OA, Beaton et al (3) reported a Cronbach's alpha of 0.87. For the 12-item WALS, a Cronbach's alpha of 0.81 has been reported in a sample of 292 patients with either OA or IA (44). Dhanhani (40) reported a Cronbach's alpha of 0.86 for an employed sample with lupus and 0.80 for those not working.

Validity. In a sample of 250 workers with either RA or OA, Beaton et al (3) reported a correlation range of r = 0.43–0.66 against a series of work-oriented constructs (self-reported global items); the WALS also showed moderate-to-high correlations against other work-specific measures, including the 6-item Stanford Presenteeism Scale (r = 0.66), Endicott Work Productivity Scale (r = 0.55), Work Instability Scale for Rheumatoid Arthritis (r = 0.77), and Work Limitations Questionnaire Index (r = 0.61).

Moreover, the WALS also showed moderate correlations against the HAQ disability index (r = 0.66), self-rated arthritis severity (r = 0.62), and self-rated pain intensity (r = 0.67), which might be considered somewhat higher than expected, given these were not work-specific indicators. None of the comparators tested in this study might be considered a gold standard reference indicator, but overall, there is moderate support for its construct validity.

Ability to detect change. In Beaton et al's sample of 250 workers with either RA or OA (3), moderate responsiveness of the 11-item WALS to 1-year improvements (standardized response mean [SRM] −0.79) and 1-year deteriorations in work ability (SRM 0.50) were found (ranked 1st out of 5 at-work measures compared in the study), but smaller effect sizes for 1-year improvements (SRM −0.37) and 1-year deterioration in work productivity (SRM 0.18) were observed (ranked 2nd out of 5 at-work measures compared). In this analysis, single item global indices of change were used to provide "reference" indicators of change, and individuals showing varying magnitudes of change were pooled in the analysis (i.e., no stratification of individuals based on magnitude of change). Additional studies to examine the responsiveness of the WALS to more defined changes (i.e., "smaller" versus "larger") in workplace activity limitations could be informative.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. This is a feasible measure that captures many of the key perspectives of workplace activity limitations, and has shown very good psychometric performance overall. Included in the scale is an assessment of potential difficulties related to mobility (around workplace, to and from work), which is often omitted in other work outcomes. It also offers a difficulty unrelated to arthritis response option (scored as 0), which is also quite unique. To date, available psychometric evidence has been gathered from samples with OA, IA, and lupus, suggesting potential utility across different arthritis conditions.

Caveats and cautions. The WALS has yet to be examined to date as a true study end point or applied in clinical trials. There is also limited current information on interpretation of scores or cut points. Users should be aware of the 2 different versions of the measure, which have shown comparable levels of internal consistency (comparability of other psychometric properties not yet known). Recall time frame is not a specified time, but is "in general" or "typically."

Clinical usability. The WALS has good potential. It is a feasible tool that can inform impact of disease at the "activity" level (specific to the workplace). When used in concert with other health indicators, it has the potential to help guide clinical decisions related to management strategies (e.g., the need for additional therapeutic or workplace interventions) and vocational recommendations and/or decisions (e.g., sick leave); establishing clinical parameters (e.g., minimum clinically important difference, patient acceptable symptomatic state) in future studies will be useful to this end.

Research usability. Research usability is good, with low respondent burden and minimal administrative burden.

WORK INSTABILITY SCALE FOR RHEUMATOID ARTHRITIS (RA-WIS)

Description

Purpose. Measures the extent of work instability (WI), which is defined as "a state in which the consequences of a mismatch between an individual's functional abilities and the demands of his or her job can threaten continuing employment if not resolved" (13). Originally developed specifically for rheumatoid arthritis (RA).

Content. This is a psychometric (summative) measure with items covering a broad range of specific work-related issues (e.g., symptom control, time management, task difficulties at work, cognitive distresses due to concerns about future employability) that may signify a functional ability/job demands mismatch.

Number of items. 23.

Response options/scale. Dichotomous response options: yes/no.

Recall period for items. "At the moment."

Endorsements. None. (The RA-WIS is 1 of 6 measures identified as a candidate Outcome Measures in Rheumatology work productivity outcome.)

Examples of use. RA-WIS was assessed as a secondary outcome in a clinical trial of anti–tumor necrosis factor adalimumab on RA (46), and as an outcome for comparing occupational therapy versus usual care in RA (47). Macedo et al (48) examined the relationships between the Disease Activity Score using 28-joint counts, Health Assessment Questionnaire (HAQ) scores, and RA-WIS (as study outcome). The RA-WIS also showed ability to predict arthritis-related work transitions (e.g., disability leave of absence, reducing work hours, or job changes) among workers with RA or osteoarthritis (OA) within 1 year (49).

Practical Application

How to obtain. The RA-WIS is copyrighted to the Psychometric Laboratory for Health Sciences, University of Leeds; further information available at URL: www.leeds.ac.uk/medicine/rehabmed/psychometric/Scales3.htm.

Method of administration. Patient administration (self-report).

Scoring. The RA-WIS is scored by summing responses from all 23 scale items (scale range 0–23), no instructions for missing value available; conversion into interval-level scaling (derived from Rasch analysis) is available for OA (50).

Score interpretation. Cut points have been established to differentiate levels of WI: low <10, moderate 10–17, and high >17 (13). This 3-level categorization has also demonstrated predictive validity for arthritis-related work transitions within 1 year (49).

Respondent burden. Low. Easy to read.

Administrative burden. Minimal.

Translations/adaptations. Available in 18 languages. Adaptation of the RA-WIS requires explicit permission from Galen Research in Manchester. Cross-cultural validity of the RA-WIS has been shown for English, Dutch, German, and French (51).

Psychometric Information

Method of development. Details of original development of the RA-WIS are reported in Gilworth et al (13). Qualitative interviews were conducted in individuals with RA to identify key themes (job flexibility, good working relationships, and symptom control) from which items were generated; 76 initial statements were identified as potential items, which were reduced to 36 items based on ability to discriminate against 5 levels of work instability as assessed by vocational experts, and then finally reduced to 23 items based on assessment of fit to the Rasch model (1-parameter item response theory approach).

Acceptability. In a sample of 250 workers with either RA or OA, Beaton et al (3) reported that RA-WIS scores were available in 223 patients (scores were not calculated if >10% of items were missed), of whom 9.4% had the floor score (RA-WIS = 0), while 0.4% were at the ceiling score (RA-WIS = 23). Overall, readability of RA-WIS items is high.

Reliability. Gilworth et al (13) reported a test–retest correlation of r = 0.89 (n = 51); in a sample of 250 workers with RA or OA (3), Kuder-Richardson Formula-20 (KR-20) for the RA-WIS was 0.91 and item-total correlation ranged from 0.34–0.71. KR-20 was 0.93 for 130 patients with OA within this sample (50).

Validity. In a sample of 250 workers with RA or OA (3), the level of correlation (Spearman's rank correlation coefficient) against work-oriented constructs (self-reported global items) was r = 0.54–0.74, which was considered best among 5 at-work measures compared in this study. The RA-WIS also showed moderate-to-high correlations against other work-specific measures, including the Workplace Activity Limitations Scale (r = 0.77), 6-item Stanford Presenteeism Scale (r = 0.69), Endicott Work Productivity Scale (r = 0.64), and Work Limitations Questionnaire Index (r = 0.61), and correlated moderately with the HAQ (r = 0.66), self-rated arthritis severity (r = 0.62), and pain intensity (r = 0.67). Among workers with OA (n = 130) (50), the range of correlation (Spearman's rank correlation coefficient) against work-oriented constructs (self-reported global items) was r = 0.55–0.77, and moderate-to-high correlations with the HAQ (r = 0.70), arthritis severity (r = 0.75), and pain intensity (r = 0.79) were also found. Of note is that the level of correlations between the RA-WIS and health (non–work specific) measures from these studies appeared to be somewhat higher than might be expected (i.e., comparable to the level of correlation against work-specific measures). Tang et al (50) demonstrated that the scoring structure of the RA-WIS shows adequate fit to the expectations of the Rasch model with only minor modifications, and therefore its summed score may be considered compatible for transformation into interval-level scaling. Proper fit to the Rasch model requires the pattern of item response to satisfy a number of criteria, including: 1) approximation to the Guttman structure, 2) demonstrating the lack of differential item functioning, as well as 3) providing evidence of unidimensionality and local independence of items. These criteria were met when RA-WIS was tested in OA.

Ability to detect change. In a sample of 250 workers with either RA or OA (3), high responsiveness of the RA-WIS to 1-year improvements (standardized response mean [SRM] −0.64) and 1-year deteriorations in work ability (SRM 0.88) were found (ranked 2nd out of 5 at-work measures compared in the study), but only small-to-negligible effect sizes for 1-year improvements (SRM −0.29) and deteriorations in work productivity (SRM 0.00) were found (ranked 5th out of 5 at-work measures compared in the study). In a sample of 130 patients with OA, large effect sizes were observed in the RA-WIS for 1-year deteriorations in intrusiveness of arthritis on work (SRM 1.05), and also for 1-year improvements in intrusiveness of arthritis on work (SRM −0.78) (50). In both of these studies, single item global ratings of change were used to provide "reference" indicators, and it should be considered that individuals demonstrating various magnitudes of change were pooled into the same analysis. Further studies to examine whether the RA-WIS is similarly responsive to more defined changes ("smaller" versus "larger") in work instability could be informative.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. "Work instability" (functional ability/job demands mismatch) is a unique concept among available work outcomes; evidence of ability to predict for future work transition outcomes in RA and OA suggests that it could have a potential role in risk prognostication, in addition to potential applications as a study end point. Patient preference for the RA-WIS over 3 other at-work disability measures was shown in a study that sampled injured workers with upper-extremity musculoskeletal disorders (4), likely due to the relative simplicity of item statements and/or the dichotomous response options. Overall psychometric evidence is very good in both RA and OA. Alternative versions of the scale have been developed for other rheumatic conditions like ankylosing spondylitis (Work Instability Scale for Ankylosing Spondylitis) (17).

Caveats and cautions. Some concerns about the limited response options (i.e., lack of a middle-ground option between "yes" and "no") have been suggested by workers with RA or OA (Beaton et al: unpublished observations). Some items appeared to exhibit some redundancy (50).

Clinical usability. The RA-WIS is feasible and is a measure that has been well received by patients with arthritis or other musculoskeletal conditions. It has potential for clinical use for risk prognostication of adverse future work outcomes. Support for the predictive validity of the proposed RA-WIS cut points has been shown, which could be applied for risk stratification (low versus moderate versus high WI). Further psychometric evidence and establishing clinical parameters (e.g., minimum clinically important difference, patient acceptable symptomatic state) will be helpful for clinical interpretation of scores.

Research usability. Research usability is excellent. It is a versatile tool (intended concept could be of interest as either a prognostic factor or study end point). There is minimal administrative and respondent burden, unless score calibrations are used (i.e., conversion of summed scores to interval-level scores).

WORK LIMITATIONS QUESTIONNAIRE (WLQ)

Description

Purpose. Measures the on-the-job impact of chronic health conditions and treatment with a focus on assessing limitations while performing specific job demands. Original 25-item version (WLQ-25) was published in 2001 (14), and several shortened versions have been tested or applied in studies in workers with various musculoskeletal disorders: WLQ-16 (32), WLQ-8 (30,31), and a 6-item version (33). Another variant is the Work Role Functioning scale (WRF-26) (22), which has similar purpose and content, but is reversed in conceptual orientation (i.e., assesses level of role functioning, not limitations). The current review will focus mainly on the WLQ-25.

Content. The WLQ-25 can be organized into 4 domains: time management (TM), addresses difficulty with handling a job's time and scheduling demands; physical demands (PD), examines ability to perform job tasks that involve bodily strength, movement, endurance, coordination, and flexibility; mental-interpersonal demands (MI), addresses cognitively demanding tasks and on-the-job social interactions; and output demands (OD), concerns reduced work productivity (14).

Number of items. 25 total scale items, divided into 4 subscales: TM (5 items), PD (6 items), MI (9 items), and OD (5 items).

Response options/scale. Three of the 4 subscales (TM, MI, OD) examine proportion of time with difficulty: "none of the time (0%)," score = 0; "a slight bit of the time," score = 1; "some of the time (50%)," score = 2; "most of the time," score = 3; "all of the time (100%)," score = 4; plus a "does not apply to my job" option (treated as missing, no score). The PD subscale has reverse instructions and examines proportion of time without difficulty (same response options provided).

Recall period for items. 2 weeks for the WLQ-25. This varies with other versions.

Endorsements. None. (The WLQ-25 is 1 of 6 measures identified as a candidate Outcome Measures in Rheumatology work productivity outcome.)

Examples of use. Rohekar and Pope (52) applied the WLQ-25 as an outcome to assess work disability in seronegative spondylarthritis. Allaire et al (53) applied the WLQ-25 to examine the impact of rheumatoid arthritis (RA) on work disability among 5,419 older workers (age range 55–64 years) with RA. Lerner et al (54) recently developed a method to impute work productivity impact using other health variables, following an examination of the relationship between the WLQ-25 and an array of pain, functioning, and general health measures. Tang et al (4) provided a head-to-head comparison of the psychometric performance of 4 at-work disability measures (including WLQ-16) among injured workers with upper-extremity disorders. Zhang et al (55) examined the comparability of methods to estimate productivity loss based on 4 different instruments (includes conversions based on the WLQ Index). Associations between medical conditions (31), arthritis (56), and health risks (57) with work limitations had been examined using the WLQ-8.

Practical Application

How to obtain. The WLQ is copyrighted: Work Limitations Questionnaire 1998, by The Health Institute, Lerner D, Amick B 3rd, GlaxoWellcome. The WLQ is provided free of charge for noncommercial applications.

Method of administration. Patient administration (self-report).

Scoring. Multiple approaches have been recommended to score the WLQ-25. Subscales are scored by multiplying the mean of subscale items by 25 (range 0–100, 100 = most limitations, response orientation needs to be reversed for the PD subscale), missing response for up to 50% of subscale items is allowable for subscale scoring. Scales are scored by multiplying the mean of all scale items by 25 (range 0–100, 100 = most limitations, response orientation needs to be reversed for the PD subscale), missing response for up to 50% of subscale items is also allowable for scale scoring (Amick BC: unpublished observations). Weighted index scoring (WLQ Index) is calculated from subscale scores using a weighted formula based on an analysis of the relationship between WLQ scores and actual employee productivity loss relative to healthy employees (58). Scores from all 4 subscales are needed to calculate the WLQ Index, computer scoring is necessary. Formulas for calculating the WLQ Index and conversion to productivity loss estimates versus healthy controls (range 0–25%) are published in a technical report available from the developers (58).

Score interpretation. Normative WLQ subscale means (SE) have been reported in a small sample of healthy workers recruited from Massachusetts (37): PD = 4.5 (1.4); TM = 7.2 (3.1); MI = 10.6 (2.7); OD = 7.2 (2.8). A 10% increase in WLQ subscale score has been proposed to equate a productivity decline of 4–5% (59).

Respondent burden. Moderate. Lerner et al (60) reported an administration time of approximately 30 minutes for workers with osteoarthritis (OA), and approximately 15 minutes for healthy controls. Some workers have expressed concern over the flipping of instructions for different sections of the questionnaire (i.e., the TM, MI, and OD subscales ask about amount of time with difficulty, while the PD subscale asks about amount of time without difficulty).

Administrative burden. Moderate. Computer scoring is required for handling/imputation of missing values, and for calculating the WLQ Index.

Translations/adaptations. Over 30 official language translations.

Psychometric Information

Method of development. Content and format of the WLQ-25 were developed from focus groups (workers with chronic conditions), cognitive interviews, and an alternate form comparison (stem/response) (14); 70 job demand-level limitation items and 7 dimensions were originally generated, which were reduced to 25 items through cognitive interviews (14).

Acceptability. Lerner et al (60) reported <1% missing data in OA, floor effect = 20.4–25.8% (subscales), and ceiling effect = 0.9–2.2% (subscales). In workers with RA, Walker et al (61) reported a missing proportion of 3.1–21.8% for individual scale items and 3.1–5.6% for subscale scores, and as a result, the WLQ Index score was unavailable in 10.1% of the sample due to missing data. In a study that recruited 250 workers with RA or OA, Beaton et al (3) reported 5.6% (WLQ Index) and 19.9–35.5% (subscales) of the sample had the floor score, while 0% (WLQ Index) and 1.2–3.3% (subscales) of the sample had the ceiling score.

Reliability. Among workers with OA, Lerner et al (60) reported the following item-to-total correlation coefficients and Cronbach's alpha for the 4 WLQ-25 subscales: PD: 0.72–0.82, $\alpha = 0.93$; TM: 0.79–0.92, $\alpha = 0.95$; MI: 0.81–0.92, $\alpha = 0.97$; OD: 0.82–0.89, $\alpha = 0.96$. In RA, Walker et al (61) reported a Cronbach's alpha range of 0.83–0.88. In a pooled sample of workers with RA or OA, Beaton et al (3) reported the following item-total correlations and Cronbach's alpha for the 4 WLQ-25 subscales: PD: 0.38–0.63, $\alpha = 0.77$; TM: 0.60–0.76, $\alpha = 0.86$; MI: 0.50–0.90, $\alpha = 0.94$; OD: 0.61–0.81, $\alpha = 0.88$.

Validity. In OA (60), each of the 4 WLQ-25 subscale scores showed a logical and statistically significant association ($P < 0.001$ in analysis of variance [ANOVA] F test) with self-reported arthritis severity (4 levels: poor to very good); specific WLQ-25 subscales also showed linear associations (subscales demonstrating $P < 0.05$ in ANOVA F test in parentheses below) with level of arthritis pain (PD, TM, MI, OD), joint stiffness (PD, TM, MI, OD), functional limitations due to arthritis (PD, TM), Short Form 12 (SF-12) physical component score (PCS; PD), self-reported work productivity (OD), and work absences (OD). In RA (61), the WLQ Index showed low-to-moderate correlations with the SF-36 mental component score (r = −0.60), SF-36 PCS (r = −0.49), fatigue visual analog scale (VAS; r = 0.50), pain VAS (r = 0.46); Health Assessment Questionnaire (HAQ; r = 0.56), HAQ-II (r = 0.54), depression score (r = 0.46) and anxiety score (r = 0.41) from the Arthritis Impact Measurement Scale, days with limited activities (r = 0.38), and days unable to work (r = 0.29). Wolfe et al (62) reported similar levels of correlation between the WLQ Index and HAQ (r = 0.57), HAQ-II (r = 0.55), modified HAQ (r = 0.55), SF-36 PCS (r = 0.50), and VAS pain (r = 0.47) when tested in a population of workers with RA. In a sample of 250 workers with either RA or OA (3), the WLQ Index showed moderate correlations (r = 0.49–0.67) against a series of work-oriented constructs (self-reported global items), and also against the HAQ (r = 0.49), arthritis severity (r = 0.42), and self-rated pain intensity (r = 0.48). In this study, the WLQ Index also showed moderate-to-high correlations against other work-specific measures, including the Workplace Activity Limitations Scale (r = 0.61), 6-item Stanford Presenteeism Scale (r = 0.63), Endicott Work Productivity Scale (r = 0.61), and Work Instability Scale for Rheumatoid Arthritis (r = 0.67). Overall, relationships between the WLQ-25 and various health- and work-related indicators were by and large in line with expectations, providing strong support for its construct validity.

Ability to detect change. In a sample of 250 workers with either RA or OA (3), levels of responsiveness of the WLQ Index to 1-year improvements (standardized response mean [SRM] −0.28) and 1-year deteriorations in work ability (SRM 0.20) were modest (ranked 5th out of 5 at-work measures compared in the study). Varying levels of responsiveness to 1-year improvements (SRM −0.64) compared to 1-year deteriorations in work productivity (SRM 0.08) were also reported (tied for 3rd out of 5 at-work measures compared). In this study, single item global ratings of change were applied as "reference" indicators, and individuals experiencing varying levels of changes were pooled into the same analysis (i.e., no stratification of individuals based on magnitude of change). Further evaluations are needed to examine level of responsiveness of the WLQ-25 against more defined magnitudes of change ("smaller" versus "larger") in work limitations.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The WLQ has a solid foundation of development, which has led to well-defined domains that have been consistently applied in various studies to assess the impact of health conditions on specific aspects of work. The breadth of potential work limitations examined is high, and should have strong relevance for many job types and health conditions. To date, this is one of the most widely used measures in the field, with studies in RA, OA, and several other musculoskeletal conditions; cumulative evidence of its construct validity is excellent. The WLQ Index can be converted to provide an estimate percentage of productivity loss, a unique property among available measures in the field, which is potentially useful to bridge measurement needs for dual clinical/economic costing purposes.

Caveats and cautions. Several variations of the measure exist in the literature (WLQ-25, WLQ-16, WLQ-8, WRF-26) and approaches to score are unique among tools and across existing studies. The different versions also vary in terms of recall period, specific wording of items, and whether specific sections (i.e., PD subscale) have reverse orientation within the full questionnaire. Care must be taken when making comparisons of results across versions. For the WLQ-25, reverse instructions for the PD domain may confuse respondents and could be a source of error. There is also a generous allowance of missing data (up to 50% of missing items may be imputed).

Clinical usability. Moderate administrative and respondent burden should be considered. More studies to establish clinically meaningful cut points (e.g., minimum clinically important difference, patient acceptable symptomatic state) could help improve clinical usability.

Research usability. Excellent. Very good-to-excellent psychometric evidence in arthritis as well as in other clinical populations. High potential for use for economic costing purposes given the time orientation of response options, and purported relationship between WLQ index and level of work productivity reported by Lerner et al (59).

WORK PRODUCTIVITY AND ACTIVITY IMPAIRMENT QUESTIONNAIRE (WPAI)

Description

Purpose. Measures the effect of health and symptom severity on work productivity and nonwork activities. Two versions are available: general health (WPAI:GH) or specific health problem (WPAI:SHP), the latter is designed such that it can be modified for any health problem by specifying the disease/condition of interest in the questions (i.e., to derive disease-specific versions of the scale). The original WPAI:GH was published in 1993 (15); approach to scoring the questionnaire has changed since the original publication.

Content. Examines the extent of absenteeism, presenteeism, and impairment in daily activities attributable to general health (WPAI:GH) or a specific health problem (WPAI:SHP).

Number of items. 6 questions (Q), each with unique response options: current employment status (Q1), number of hours missed due to health problem (Q2), number of hours missed due to other reasons (Q3), hours actually worked (Q4), degree to which health affected productivity while working (Q5), degree to which health affected regular (nonwork) activities (Q6). Items are not intended to be summative.

Response options/scale. Q1: yes/no, Q2–Q4: number of hours (count data), Q5: global rating scale, 0–10 (0 = health problems had no effect on my work, 10 = health problems completely prevented me from working), Q6: global rating scale, 0–10 (0 = health problems had no effect on my daily activities, 10 = health problems completely prevented me from doing my daily activities).

Recall period for items. 7 days.

Endorsements. None. (The WPAI:GH is 1 of 6 measures identified as a candidate Outcome Measures in Rheumatology work productivity outcome.)

Examples of use. Stockl et al (63) applied the WPAI:GH as a secondary outcome to evaluate the effect of a rheumatoid arthritis (RA) disease treatment management program. In a study by Zhang et al (55), productivity costs associated with arthritis estimated with the WPAI:GH was compared against estimates made from 3 other measures (Health and Labor Questionnaire [HLQ], The Health and Work Performance Questionnaire, and Work Limitations Questionnaire-25). Haibel et al (64) applied the WPAI for ankylosing spondylitis (AS; WPAI:SpA) as an outcome measure to examine the efficacy of infliximab therapy in nonsteroidal antiinflammatory drug–refractory patients with AS. Associations among demographic, health-related, and treatment factors with the level of work productivity (WPAI:SpA) have been recently examined in patients with AS (65).

Practical Application

How to obtain. The WPAI is available at URL: www.reillyassociates.net/Index.html. Permission and fees are not required to use the WPAI.

Method of administration. Self-administered or interviewer administered.

Scoring. Detailed information is provided in the above web site; 4 outcome (OC) scores can be derived from the WPAI: OC1, percent work time missed due to health, Q2/(Q2 + Q4) (percentage of absenteeism); OC2, percent impairment while working due to health, Q5/10 (percentage of presenteeism); OC3, percent overall work impairment due to health, Q2/(Q2 + Q4) + [(1 − Q2/(Q2+Q4)) × (Q5/10)]; OC4, percent activity impairment due to health, Q6/10.

Score interpretation. For all 4 outcomes, greater scores (range 0–100%) indicate greater impact of health, clinically important cut points not yet established.

Respondent burden. Minimal.

Administrative burden. Low, only basic calculations needed.

Translations/adaptations. Translated in more than 80 languages (see URL: www.reillyassociates.net/WPAI_Translations.html). The WPAI:SHP can theoretically be adapted to any specific disease or health problem. Psychometric evidence is available for a wide range of diseases. Recently, a version of WPAI for AS has been developed (WPAI:SpA) (16).

Psychometric Information

Method of development. WPAI items were generated from 3 sources (15): 1) review of work productivity literature, 2) comments from patients with allergic rhinitis on an interviewer-administered version of the WPAI items from a series of clinical studies, and 3) cognitive debriefing of subjects following interviewer administration and self-administration of WPAI items to determine final wording.

Acceptability. In the original study that included a sample of workers with RA or other musculoskeletal disorders (15), up to 21% of the sample had missing data on WPAI questions when the measure was self-administered, but minimal missing data when the questions were interviewer administered. The WPAI:SpA had <10% missing when self-administered (16).

Reliability. In the original study by Reilly et al (15), only a modest range of Pearson's correlation coefficient (0.71–0.75) was found in a test–retest comparison of WPAI:GH fielded in workers with nonspecific health problems within the same day (at least 4 hours later). Level of agreement (e.g., intraclass correlation coefficients) between test–retest scores was not reported in this study.

Validity. Some support for the construct validity of the WPAI:GH in RA was provided in Zhang et al (66), although the scale appeared to correlate more strongly to health status indictors than work-specific comparators where the opposite might be expected. OC1 showed a correlation of r = 0.56 with the number of absent workdays in the past 3 months (question adapted from the Productivity and Disease Questionnaire), OC2 showed a correlation of r = 0.39

with the number of hours lost due to presenteeism (item from the HLQ), and OC4 showed a correlation of r = 0.39 with the number of hours getting help on unpaid work activities (item from the HLQ). In this study, 3 of 4 WPAI:GH outcomes (OC2, OC3, and OC4) also showed moderate-to-high correlations (r = 0.67–0.77) against a series of health status outcomes (function, pain, patient global estimate on health impact, fatigue, and patient global assessment of disease activity). In an adalimumab versus placebo clinical trial in AS (16), WPAI:SpA outcomes were shown to be able to discriminate between "higher" and "lower" scores (split on the basis of median score in the sample) in the Bath Ankylosing Spondylitis Disability Activity Index (BASDAI), the Ankylosing Spondylitis Quality of Life (ASQOL), Short Form 36 (SF-36) physical component summary, SF-36 mental component summary, and Health Utility Index-3 at $P < 0.05$. Evidence of known group differences is consistent with a priori expectations in the trial, although it should be recognized that the comparators applied were not work-specific indicators; additional studies to specifically examine the responsiveness of the WPAI to "true" changes in work productivity would be informative.

Ability to detect change. There are no known reports on the responsiveness of the WPAI against comparable indicators of work productivity to date, but effect sizes for persons with AS in a 24-week adalimumab versus placebo trial have been reported by Reilly et al (16). In this study, among clinical responders based on improvements in the BASDAI (>1.96 decrease in score, i.e., a large change in disease activity), standardized response means (SRMs) for WPAI outcomes were OC1: −0.25; OC2: −0.86; OC3: −0.89; OC4: −1.29. Among nonresponders (i.e., small change, no changes, or deterioration), SRMs were OC1: −0.14; OC2: −0.52; OC3: −0.54; OC4: −0.39. Among clinical responders based on changes in the ASQOL (>1.8 decrease in score, i.e., a large change in quality of life), SRMs were OC1: −0.31; OC2: −0.89; OC3: −0.94; OC4: −1.18. Among ASQOL clinical nonresponders (i.e., small change, no change, or deterioration), SRMs were OC1: −0.11; OC2: −0.46; OC3: −0.38; OC4: −0.40. It is important to exercise care when interpreting SRMs reported for "nonresponders" as this is derived from a pool of individuals who have undergone varying degrees of change over the course of the trial.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The WPAI is designed to have generalizability to a broad range of occupations/diseases, and evidence of reliability and validity is available in many musculoskeletal and nonmusculoskeletal conditions. Item content is highly consistent across different versions of the measure (WPAI:GH versus WPAI:SHP), which should facilitate comparison of outcome scores in different studies (only disease attribution of items differ between versions). It has an intuitive method to score, is compatible with economic costing (orientation of response is based on amount of time affected), and has low respondent burden.

Caveats and cautions. Four separate outcome scores are derived from the questionnaire (not intended to be summated/combined). There have been important changes in the approach to score the WPAI questionnaire since its original development.

Clinical usability. The WPAI has good potential. Establishing clinical parameters (e.g., minimum clinically important difference, patient acceptable symptomatic state) will be helpful for clinical interpretability. Administrative and respondent burdens are low.

Research usability. Research usability is good. Applicability and good psychometric performance of the WPAI has been shown in several clinical trials with patients with AS.

DISCUSSION

The current review has revealed a moderate level of evidence to date to support the psychometric properties of 5 selected measures of work disability and productivity in arthritis/musculoskeletal populations. It is important to recognize that the Work Productivity and Activity Impairment Questionnaire and Work Limitations Questionnaire were developed as generic (not disease specific) instruments, and there exists additional evidence in the literature to support their psychometric performance in other populations that have not been reviewed in detail in this article. On the other hand, evidence of psychometric performance for the Workplace Activity Limitations Scale, Work Instability Scale for Rheumatoid Arthritis, and in particular the Rheumatoid Arthritis Specific Work Productivity Survey has only emerged in the past few years, since these are relatively new measures.

We believe there is room for continued growth in this area of research. Specific measurement attributes requiring further research include examinations of test–retest reliability, validation against work-specific constructs and/or productivity data (where relevant), and assessments of responsiveness to more defined magnitudes of change (e.g., sensitivity to "smaller" versus "larger" known changes). Also, data on clinically relevant parameters (e.g., minimum clinically important difference) are scarce to date and will be important to establish to help evaluate treatment efficacy in research trials and longitudinal observational studies.

Issues on the interpretability of scores derived from measures of work disability and productivity are also of emerging interest. An important concept to recognize is that the extent of disease impact on work is ultimately a function of both the person and his or her work context (i.e., environmental factors) and the manner in which they interact (2,67). At the individual level, a change in score could reflect a change in the person's capacity to work, and/or a change in the demands of the job, for example, in the case where a work transition (e.g., job modifications, reduced work hours) has taken place to allow a person with arthritis to function better at work. To provide a more complete understanding of the bases of change over time, users may consider fielding additional instruments that can offer insights into the work context (e.g., job type, work status, contractual hours, availability of workplace support) to supplement outcome measures designed to quantify the level of work disability and productivity.

Overall, the current diversity of available measures in this field is impressive. While the availability of a wide range of instruments can provide users with many options, some care is important when selecting an outcome to meet the needs of a particular research study or clinical purpose. The specific work-related concept or measurement perspective being sought, the availability of supporting psychometric evidence, and pragmatic considerations (e.g., applicability, feasibility) should be concurrently considered. In addition to the summary of evidence provided in the current article, users may also consider additional findings and insights from a number of recent studies (3,4) that have examined the head-to-head psychometric performance of multiple work measures in arthritis/musculoskeletal populations to help inform the selection of outcomes in future research or clinical applications.

ACKNOWLEDGMENTS

The authors would like to thank members of the Outcome Measures in Rheumatology (OMERACT) Work Productivity Special Interest Group, Canadian Arthritis Network Work Productivity Group, and I-CAN-Work Alliance for contributions to this work. We also acknowledge Quenby Mahood and Denise Linton for providing assistance with literature search and manuscript preparation.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Escorpizo R, Bombardier C, Boonen A, Hazes JM, Lacaille D, Strand V, et al. Worker productivity outcome measures in arthritis. J Rheumatol 2007;34:1372–80.
2. Beaton D, Bombardier C, Escorpizo R, Zhang W, Lacaille D, Boonen A, et al. Measuring worker productivity: frameworks and measures. J Rheumatol 2009;36:2100–9.
3. Beaton DE, Tang K, Gignac MA, Lacaille D, Badley EM, Anis AH, et al. Reliability, validity, and responsiveness of five at-work productivity measures in patients with rheumatoid arthritis or osteoarthritis. Arthritis Care Res (Hoboken) 2010;62:28–37.
4. Tang K, Pitts S, Solway S, Beaton D. Comparison of the psychometric properties of four at-work disability measures in workers with shoulder or elbow disorders. J Occup Rehabil 2009;19:142–54.
5. Lavigne JE, Phelps CE, Mushlin A, Lednar WM. Reductions in individual work productivity associated with type 2 diabetes mellitus. Pharmacoeconomics 2003;21:1123–34.
6. Turpin RS, Ozminkowski RJ, Sharda CE, Collins JJ, Berger ML, Billotti GM, et al. Reliability and validity of the Stanford Presenteeism Scale. J Occup Environ Med 2004;46:1123–33.
7. Ozminkowski RJ, Goetzel RZ, Chang S, Long S. The application of two health and productivity instruments at a large employer. J Occup Environ Med 2004;46:635–48.
8. Meerding WJ, IJzelenberg W, Koopmanschap MA, Severens JL, Burdorf A. Health problems lead to considerable productivity loss at work among workers with high physical load jobs. J Clin Epidemiol 2005; 58:517–23.
9. Brouwer WB, Koopmanschap MA, Rutten FF. Productivity losses without absence: measurement validation and empirical evidence. Health Policy 1999;48:13–27.
10. Sanderson K, Tilse E, Nicholson J, Oldenburg B, Graves N. Which presenteeism measures are more sensitive to depression and anxiety? J Affect Disord 2007;101:65–74.
11. Osterhaus JT, Purcaru O, Richard L. Discriminant validity, responsiveness and reliability of the rheumatoid arthritis-specific Work Productivity Survey (WPS-RA). Arthritis Res Ther 2009;11:R73.
12. Gignac MA, Badley EM, Lacaille D, Cott CC, Adam P, Anis AH. Managing arthritis and employment: making arthritis-related work changes as a means of adaptation. Arthritis Rheum 2004;51:909–16.
13. Gilworth G, Chamberlain MA, Harvey A, Woodhouse A, Smith J, Smyth MG, et al. Development of a work instability scale for rheumatoid arthritis. Arthritis Rheum 2003;49:349–54.
14. Lerner D, Amick BC 3rd, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. Med Care 2001;39:72–85.
15. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993;4:353–65.
16. Reilly MC, Gooch KL, Wong RL, Kupper H, van der Heijde D. Validity, reliability and responsiveness of the Work Productivity and Activity Impairment Questionnaire in ankylosing spondylitis. Rheumatology (Oxford) 2010;49:812–9.
17. Gilworth G, Emery P, Barkham N, Smyth MG, Helliwell P, Tennant A. Reducing work disability in ankylosing spondylitis: development of a work instability scale for AS. BMC Musculoskelet Disord 2009;10:68.
18. Osterhaus JT, Gutterman DL, Plachetka JR. Healthcare resource and lost labour costs of migraine headache in the US. Pharmacoeconomics 1992;2:67–76.
19. Van Roijen L, Essink-Bot ML, Koopmanschap MA, Bonsel G, Rutten FF. Labor and health status in economic evaluation of health care: the Health and Labor Questionnaire. Int J Technol Assess Health Care 1996;12:405–15.
20. Endicott J, Nee J. Endicott Work Productivity Scale (EWPS): a new measure to assess treatment effects. Psychopharmacol Bull 1997;33: 13–6.
21. Kopec JA, Esdaile JM. Occupational role performance in persons with back pain. Disabil Rehabil 1998;20:373–9.
22. Amick BC 3rd, Lerner D, Rogers WH, Rooney T, Katz JN. A review of health-related work outcome measures and their uses, and recommended measures. Spine (Phila Pa 1976) 2000;25:3152–60.
23. Altshuler L, Mintz J, Leight K. The Life Functioning Questionnaire (LFQ): a brief, gender-neutral scale assessing functional outcome. Psychiatry Res 2002;112:161–82.
24. Koopman C, Pelletier KR, Murray JF, Sharda CE, Berger ML, Turpin RS, et al. Stanford presenteeism scale: health status and employee productivity. J Occup Environ Med 2002;44:14–20.
25. Kumar RN, Hass SL, Li JZ, Nickens DJ, Daenzer CL, Wathen LK. Validation of the Health-Related Productivity Questionnaire Diary (HRPQ-D) on a sample of patients with infectious mononucleosis: results from a phase 1 multicenter clinical trial. J Occup Environ Med 2003;45:899–907.
26. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). J Occup Environ Med 2003;45:156–74.
27. Goetzel RZ, Ozminkowski RJ, Long SR. Development and reliability analysis of the Work Productivity Short Inventory (WPSI) instrument measuring employee health and productivity. J Occup Environ Med 2003;45:743–62.
28. Shikiar R, Halpern MT, Rentz AM, Khan ZM. Development of the Health and Work Questionnaire (HWQ): an instrument for assessing workplace productivity in relation to worker health. Work 2004;22: 219–29.
29. Stewart WF, Ricci JA, Leotta C, Chee E. Validation of the work and health interview. Pharmacoeconomics 2004;22:1127–40.
30. Burton WN, Chen CY, Conti DJ, Pransky G, Edington DW. Caregiving for ill dependents and its association with employee health risks and productivity. J Occup Environ Med 2004;46:1048–56.
31. Burton WN, Pransky G, Conti DJ, Chen CY, Edington DW. The association of medical conditions and presenteeism. J Occup Environ Med 2004;46 Suppl:S38–45.
32. Beaton DE, Kennedy CA. Beyond return to work: testing a measure of at-work disability in workers with musculoskeletal pain. Qual Life Res 2005;14:1869–79.
33. Munir F, Jones D, Leka S, Griffiths A. Work limitations and employer adjustments for employees with chronic illness. Int J Rehabil Res 2005;28:111–7.
34. Feuerstein M, Nicholas RA, Huang GD, Haufler AJ, Pransky G, Robertson M. Workstyle: development of a measure of response to work in those with upper extremity pain. J Occup Rehabil 2005;15:87–104.
35. Feuerstein M, Nicholas RA. Development of a short form of the Workstyle measure. Occup Med (Lond) 2006;56:94–9.
36. Van Roijen L. Short Form - Health and Labour Questionnaire. Institute for Medical Technology Assessment, Erasmus MC, University Medical Centre Rotterdam; 2007.
37. Hazes JM, Taylor P, Strand V, Purcaru O, Coteur G, Mease P. Physical function improvements and relief from fatigue and pain are associated with increased productivity at work and at home in rheumatoid arthri-

tis patients treated with certolizumab pegol. Rheumatology (Oxford) 2010;49:1900–10.

38. Kavanaugh A, Smolen JS, Emery P, Purcaru O, Keystone E, Richard L, et al. Effect of certolizumab pegol with methotrexate on home and work place productivity and social activities in patients with active rheumatoid arthritis. Arthritis Rheum 2009;61:1592–600.
39. Gignac MA, Cao X, Lacaille D, Anis AH, Badley EM. Arthritis-related work transitions: a prospective analysis of reported productivity losses, work changes, and leaving the labor force. Arthritis Rheum 2008;59: 1805–13.
40. Dhanhani A. The workplace challenges of patients with lupus. Graduate Department of the Institute of Medical Science, University of Toronto; 2010. URL: http://hdl.handle.net/1807/25407.
41. Gignac MA, Cao X, Tang K, Beaton DE. An examination of arthritis-related work place activity limitations and intermittent disability over four-and-a-half years and its relationship to job modifications and outcomes. Arthritis Care Res (Hoboken) 2011;63:953–62.
42. Gignac MA. Arthritis and employment: an examination of behavioral coping efforts to manage workplace activity limitations. Arthritis Rheum 2005;53:328–36.
43. Gignac MA, Sutton D, Badley EM. Reexamining the arthritis-employment interface: perceptions of arthritis-work spillover among employed adults. Arthritis Rheum 2006;55:233–40.
44. Gignac MA, Sutton D, Badley EM. Arthritis symptoms, the work environment, and the future: measuring perceived job strain among employed persons with arthritis. Arthritis Rheum 2007;57:738–47.
45. Gignac MA, Cao X. “Should I tell my employer and coworkers I have arthritis?” A longitudinal examination of self-disclosure in the work place. Arthritis Rheum 2009;61:1753–61.
46. Bejarano V, Quinn M, Conaghan PG, Reece R, Keenan AM, Walker D, et al. Effect of the early use of the anti–tumor necrosis factor adalimumab on the prevention of job loss in patients with early rheumatoid arthritis. Arthritis Rheum 2008;59:1467–74.
47. Macedo AM, Oakley SP, Panayi GS, Kirkham BW. Functional and work outcomes improve in patients with rheumatoid arthritis who receive targeted, comprehensive occupational therapy. Arthritis Rheum 2009; 61:1522–30.
48. Macedo A, Oakley S, Gullick N, Kirkham B. An examination of work instability, functional impairment, and disease activity in employed patients with rheumatoid arthritis. J Rheumatol 2009;36:225–30.
49. Tang K, Beaton DE, Gignac MA, Lacaille D, Zhang W, Bombardier C, and the Canadian Arthritis Network Work Productivity Group. The Work Instability Scale for rheumatoid arthritis predicts arthritis-related work transitions within 12 months. Arthritis Care Res (Hoboken) 2010; 62:1578–87.
50. Tang K, Beaton DE, Lacaille D, Gignac MA, Zhang W, Anis AH, et al. The Work Instability Scale for Rheumatoid Arthritis (RA-WIS): does it work in osteoarthritis? Qual Life Res 2010;19:1057–68.
51. Gilworth G, Emery P, Gossec L, Vliet Vlieland TP, Breedveld FC, Hueber AJ, et al. Adaptation and cross-cultural validation of the rheumatoid arthritis work instability scale (RA-WIS). Ann Rheum Dis 2009; 68:1686–90.
52. Rohekar S, Pope J. Assessment of work disability in seronegative spondyloarthritis. Clin Exp Rheumatol 2010;28:35–40.
53. Allaire S, Wolfe F, Niu J, Lavalley M, Michaud K. Work disability and its economic effect on 55–64-year-old adults with rheumatoid arthritis. Arthritis Rheum 2005;53:603–8.
54. Lerner D, Chang H, Rogers WH, Benson C, Schein J, Allaire S. A method for imputing the impact of health problems on at-work performance and productivity from available health data. J Occup Environ Med 2009;51:515–24.
55. Zhang W, Gignac MA, Beaton D, Tang K, Anis AH. Productivity loss due to presenteeism among patients with arthritis: estimates from 4 instruments. J Rheumatol 2010;37:1805–14.
56. Burton WN, Chen CY, Schultz AB, Conti DJ, Pransky G, Edington DW. Worker productivity loss associated with arthritis. Dis Manag 2006;9: 131–43.
57. Burton WN, Chen CY, Conti DJ, Schultz AB, Pransky G, Edington DW. The association of health risks with on-the-job productivity. J Occup Environ Med 2005;47:769–77.
58. Lerner D, Rogers WH, Chang H. Scoring the work limitations questionnaire (WLQ) and the WLQ index for estimating work productivity loss. Technical Report. Boston: The Health Institute, Tufts-New England Medical Center; 2003.
59. Lerner D, Amick BC 3rd, Lee JC, Rooney T, Rogers WH, Chang H, et al. Relationship of employee-reported work limitations to work productivity. Med Care 2003;41:649–59.
60. Lerner D, Reed JI, Massarotti E, Wester LM, Burke TA. The Work Limitations Questionnaire’s validity and reliability among patients with osteoarthritis. J Clin Epidemiol 2002;55:197–208.
61. Walker N, Michaud K, Wolfe F. Work limitations among working persons with rheumatoid arthritis: results, reliability, and validity of the work limitations questionnaire in 836 patients. J Rheumatol 2005; 32:1006–12.
62. Wolfe F, Michaud K, Choi HK, Williams R. Household income and earnings losses among 6,396 persons with rheumatoid arthritis. J Rheumatol 2005;32:1875–83.
63. Stockl KM, Shin JS, Lew HC, Zakharyan A, Harada AS, Solow BK, et al. Outcomes of a rheumatoid arthritis disease therapy management program focusing on medication adherence. J Manag Care Pharm 2010;16: 593–604.
64. Haibel H, Song IH, Rudwaleit M, Listing J, Hildemann S, Sieper J. Multicenter open-label study with infliximab in active ankylosing spondylitis over 28 weeks in daily practice. Clin Exp Rheumatol 2008; 26:247–52.
65. Maksymowych WP, Gooch KL, Wong RL, Kupper H, van der Heijde D. Impact of age, sex, physical function, health-related quality of life, and treatment with adalimumab on work status and work productivity of patients with ankylosing spondylitis. J Rheumatol 2010;37:385–92.
66. Zhang W, Bansback N, Boonen A, Young A, Singh A, Anis AH. Validity of the work productivity and activity impairment questionnaire: general health version in patients with rheumatoid arthritis. Arthritis Res Ther 2010;12:R177.
67. Sandqvist JL, Henriksson CM. Work functioning: a conceptual framework. Work 2004;23:147–57.

Summary Table for Work Disability and Productivity Measures*

Measure	Purpose/ content	Method of administration	Respondent burden	Administration burden	Score interpretation	Reliability evidence†	Validity evidence†	Ability to detect change†	Strengths	Cautions
Rheumatoid Arthritis Specific Work Productivity Survey (WPS-RA)	Impact of RA on employment work, household work, daily activities	Interviewer administered	Low	Low	Q1: n/a; Q2, Q3,Q5–Q8: number of days affected; Q4, Q9 (0 = no interference, 10 = complete interference)	TRT: not yet established; IC: n/a	Modest (RA)	Modest (RA)	Considers both absenteeism and presenteeism; considers impacts on both employment and household work	Eight separate outcome scores are derived (not intended to be summative)
Workplace Activity Limitations Scale (WALS)	Physical functioning, activity limitations at work	Self- or interviewer administered	Low	Minimal	Range 0–33 (11 item WALS) or 0–36 (12 item WALS), higher score means more workplace activity limitations	TRT: not yet established; IC: good (OA, IA)	Very good (OA, RA)	Very good (OA, RA)	Brief tool that examines key aspects of workplace activity limitations; comprehensive response key which includes “not applicable” and “difficulties unrelated to arthritis” options	Recall time frame is not specified, but is “in general” or “typically”
Work Instability Scale for Rheumatoid Arthritis (RA-WIS)	WI: degree of mismatch between functional abilities and job demands	Self-administered	Low	Minimal to low (interval-level score transformation available)	Range 0–23, 23 = highest WI; proposed cut points: <10 = low WI, 10–17 = moderate WI, >17 = high WI	TRT: excellent (RA); IC: excellent (RA, OA)	Very good (RA, OA)	Very good (RA, OA)	Is a unique concept among available work outcome measures in the field; patient preference for this scale shown in arthritis/musculoskeletal populations	Dichotomous response options may be too coarse; some items appeared redundant
Work Limitations Questionnaire (WLQ)	Measures proportion of time with difficulty performing various aspects of work	Self-administered	Moderate	Moderate	Range 0–100 (4 subscales), 100 = most limitations; 10% increase in subscale score equivalent to 4–5% loss in work productivity	IC: very good (RA, OA)	Excellent (RA, OA)	Good (RA, OA)	Well-established domain structure; validated across different musculoskeletal disease and chronic health populations; WLQ Index can be converted to provide an estimate of % productivity loss (against healthy controls)	Multiple variants of WLQ and scoring approaches exist, caution is required when comparing findings; reversed instructions for the physical demands subscale vs other subscales
Work Productivity and Activity Impairment Questionnaire (WPAI)	Extent of absenteeism, presenteeism, and impairments on daily activities	Self- or interviewer administered	Minimal	Low	4 different outcome scores (each with range 0–100, indicating % impairment)	TRT: modest (RA); IC: n/a	Modest (RA, AS); additional evidence available in other populations (not appraised here)	Good (RA, AS)	Evidence of reliability and validity has been shown across many disease populations; low respondent burden; high potential for economic costing applications	Four outcome scores are derived; examines global impact (difficulties with specific aspects of work not assessed)

* RA = rheumatoid arthritis; Q = question; n/a = not applicable; TRT = test–retest; IC = internal consistency; OA = osteoarthritis; IA = inflammatory arthritis; WI = work Instability; AS = ankylosing spondylitis.

† Specific arthritis/rheumatic population indicated in parentheses where evidence is available.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S59–S63
DOI 10.1002/acr.20622
© 2011, American College of Rheumatology

Gout Measures

Gout Assessment Questionnaire (GAQ, GAQ2.0), and Physical Measurement of Tophi

WILLIAM J. TAYLOR

INTRODUCTION

Outcomes assessment of gout has been a relatively neglected area of rheumatology measurement science until recent years, in which the paucity of properly tested instruments have been highlighted by clinical trials of therapy for acute gout attacks and more recently preventative treatment of chronic gout. This paucity may have been due to the traditional reliance on simple pain responses for acute gout and upon serum urate changes for chronic gout as the typical outcomes of interest. This has changed significantly since gout became a topic for the Outcome Measures in Rheumatology Clinical Trials (OMERACT) in 2004 (1) and the general recognition of patient reported outcomes as vital for proper understanding of the effect of treatment.

This review is based upon work conducted through the OMERACT process, review of the Ovid Medline database (to August 2010) concerning the keywords “gout” AND [“outcome measure.mp” OR “Questionnaire”], personal archives of the author, and other work conducted by the author in collaboration with colleagues from the OMERACT Gout Working Group.

GOUT ASSESSMENT QUESTIONNAIRE (GAQ, GAQ2.0)

Description

Purpose. The original GAQ, reported in 2006 (2), was developed to fill a large gap, there being no other gout-specific patient reported outcome instrument. It was conceived as measuring the impact of gout and its treatment from the patient’s perspective, but was developed largely within the context of a single clinical trial. The GAQ2.0, reported in 2008 was developed with more patient involvement and was tested in a community-based sample of gout patients (3).

Content. The GAQ is a 21-item questionnaire that collects information about gout impact, assessing pain, well-being, productivity, and treatment satisfaction. The GAQ2.0 contains a Gout Impact Scale (GIS) and 4 other sections that collect clinical, background, and economic data that are not scored. The GIS scores the domains of overall concern, medication side effects, perception of unmet needs, and impact of acute episodes.

Number of items. There are 21 items in the GAQ. There are a total of 31 questions in the GAQ2.0, but most of these are categorical or designed to be reported as individual items, and are not summated. There are 24 items (in 3 questions) in the GIS portion of the GAQ2.0, which are summated to form 5 scales. The other questions describe the respondents’ gout, recent attacks, treatment, medical history, and demographics. These additional questions are not formally scored.

Response options/scale. Most items of the GAQ are scored on a Likert scale and some are scored by number of days of hours of activity restriction. There are 5 sections of the GAQ2.0 with 31 items. Each item has different response options. Each item of the GIS portion of the GAQ2.0 is rated “strongly agree” to “strongly disagree,” “all of the time” to “none of the time,” or “not a bit” to “extremely” on a 5-point Likert scale.

Recall period for items. There is no stated recall period.

Endorsements. The instrument has not been endorsed by any group.

Examples of use. The instrument has not been reported in any published study, except for the original 2 development studies. The instrument developers have published other articles, but these are presentations of data from the same group of patients studied in the instrument development process. One article focused on the Short Form 36, version 2 (SF-36) scores and categories of gout characteristics from the “Gout Background Questionnaire” (presumably a component of the GAQ2.0, although this was not explicitly stated) (4). Another article focused on health care utilization (5) and another focused on discrepancies between patient and physician rating of gout severity (6).

William J. Taylor, PhD, MBChB: University of Otago, Wellington, and Hutt Valley District Health Board, Lower Hutt, New Zealand.

William J. Taylor, PhD, MBChB, Rehabilitation Teaching and Research Unit, University of Otago Wellington, PO Box 7343, Wellington, New Zealand. E-mail: will.taylor@otago.ac.nz.

Submitted for publication January 21, 2011; accepted in revised form April 12, 2011.

Practical Application

How to obtain. The instruments were developed through a pharmaceutical development program by TAP Pharmaceutical Products (now Takeda Pharmaceuticals), which retains copyright. However, GAQ2.0 is freely available from Dr. Omar Dabbous, Senior Director of Global Health Economics and Outcomes, Takeda Pharmaceuticals International Deerfield, IL (E-mail: omar.dabbous@tpna.com).

Method of administration. The instrument is self-reported.

Scoring. The GAQ is scored in 7 subscales: gout concern, well-being, productivity, gout pain and severity, treatment convenience, treatment satisfaction, and treatment bother. Each subscale contains 1–6 items. Subscales are reported in a 0–100 range but the detailed scoring procedure is not reported.

The GIS portion of the GAQ2.0 is scored in 5 subscales (total of 24 items): gout concern overall (4 items), gout medication side effects (2 items), unmet gout treatment need (3 items), well-being during attack (11 items), and gout concern during attack (4 items). Each item is scored on a 5-point Likert scale. Subscales are reported in a 0–100 range but the detailed scoring procedure is not reported.

Score interpretation. Higher scores indicate more problems.

Respondent burden. The time needed to complete the GAQ or GAQ2.0 has not been reported. The GAQ2.0 consists of 7 pages and the GIS portion is 1.5 pages.

Administrative burden. Not reported.

Translations/adaptations. There are no language or cultural translations. The instrument was developed in the US in 3 centers, mainly with male subjects.

Psychometric Information

Method of development. Items for the original GAQ were identified mainly through literature review. Items were potentially modified through telephone interview with 5 gout patients after the draft questionnaire was completed by postal survey. The instrument was tested in a phase 2 clinical trial of febuxostat compared to placebo (126 patients). The subscales were formed through factor analysis but the details of this analysis have not been published.

During development of the GAQ2.0, 2 focus groups were conducted but the method of qualitative analysis was unclear. Some new items were added as a result of the focus group interviews. The GAQ2.0 was tested in a community cohort of patients with gout (297 people) and analysed using Rasch modelling and confirmatory factor analysis with structural equation modelling.

Acceptability. Not reported.

Reliability. The GAQ instrument was not evaluated for test–retest reliability. Approximately one-fifth of the validation sample completed the GAQ2.0 on 2 occasions over 2 weeks. The intraclass correlation coefficients (ICC) ranged from 0.77–0.89 for the 5 subscales of the GIS, but it was not clear which subscale belonged to which ICC.

Validity. The internal consistency of the GAQ subscales was assessed using Cronbach's alpha, and ranged from 0.83–0.97. This statistic was not suitable for the single item scales (treatment convenience, treatment satisfaction). Construct validity was assessed by correlation with the SF-36 subscales. Correlation was generally low (<0.45) for each GAQ subscale (7). The highest correlations for each subscale were well-being (0.30 SF-36 bodily pain), productivity (0.30 SF-36 role-physical), gout concern (0.41 SF-36 bodily pain), treatment satisfaction (0.35 SF-36 bodily pain), gout pain and severity (0.45 SF-36 bodily pain), treatment bother (0.20 SF-36 vitality), and treatment convenience (0.14 SF-36 vitality).

Internal consistency assessed using Cronbach's alpha of the GAQ2.0 GIS scales ranged from 0.60–0.94. The 2-item Gout Medication Side Effects scale and the 3-item Unmet Gout Treatment Need scale had poor internal consistency (0.60 and 0.65, respectively.) Although item-fit statistics were presented for the Rasch analysis, overall model fit, formal tests for unidimensionality, local dependence, and item bias were not reported.

The construct validity of the GAQ2.0 GIS scales is difficult to discern since the scales are rather idiosyncratic and are poorly represented by any other reported scale or concept. However, all reported correlations are low in magnitude and some of these are not supportive of construct validity. In particular, the correlation between gout concern overall and patient-rated severity was only 0.45; unmet gout treatment need and attack frequency in the past year was 0.43; gout concern during an attack and typical attack pain during the past 3 months was 0.21; the physical functioning scale and general health scale of the SF-36 version 2 failed to correlate beyond 0.3 with any of the GIS scales.

Ability to detect change. There are no data to show that GIS scales change over time.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The GAQ2.0 is the only published gout-specific instrument that attempts to measure the impact of gout from the perspective of the patient, and to comprehensively describe the experience of having gout.

Caveats and cautions. Two subscales of the GIS portion of the GAQ2.0 were considered by OMERACT 10 and were not endorsed as having sufficiently met the OMERACT filter for use in clinical trials of chronic gout (8). These subscales were gout concern overall scale and unmet need scale. The construct validity of all 5 scales of the GIS portion of the GAQ2.0 is unclear. The overall concept of "impact of disease" is ambiguous and not well-defined.

Clinical usability. The instrument is not recommended for routine clinical use at this time.

Research usability. The instrument is not recommended for use in research settings at this time, except where the purpose of the research is further refinement of the instrument.

Table 1. Definitions for change in tophus burden using digital photography (10)

Measurement	Tophus response	Patient response*
For ≤5 measurable tophi		
100% decrease in tophus area	Complete response	Complete response
≥75% decrease in tophus area	Marked response	Partial response
≥50% decrease in tophus area	Partial response	Partial response
Neither a 50% decrease nor 25% decrease in tophus area	Stable disease	Stable disease
≥25% increase in the tophus area	Progressive disease	Progressive disease
For ≤2 nonmeasurable tophi		
Disappearance of the tophi	Complete response	Complete response
Approximately ≥50% reduction in size	Improved	Partial response
Neither improvement nor progression can be determined	Stable disease	Stable disease
Approximately ≥50% increase in the area of the tophus	Progressive disease	Progressive disease

* Defined as the best tophus response in the absence of a new tophus or progressive disease in any tophus (in which case the response is progressive disease).

PHYSICAL MEASUREMENT OF TOPHI (TAPE MEASUREMENT, VERNIER CALIPERS, ENUMERATION, DIGITAL PHOTOGRAPHY)

Description

Purpose. Tophi are pathognomonic of chronic gout and may be responsible for joint damage, as well as being unsightly and intrinsically undesirable. Tophi are a legitimate target for treatment (9) and therefore require a satisfactory method of measurement. A number of physical methods have been used to achieve this purpose and will be discussed here. The purpose of these techniques is primarily to determine response to therapy that might reduce tophus burden.

Content. Enumeration of tophi by simply counting the total number of palpable tophi is a rapid and inexpensive method. The tape measurement technique has been described to determine the area of a sentinel tophus and uses a standard tape measure to identify the distance between 2 pen marks drawn on a predefined length and width axis that are orthogonal to each other. The area is calculated as the product of these 2 distances. Vernier calipers (150 mm digital) have also been used to determine the longest diameter of a sentinel tophus. Digital photography using a standardized image acquisition protocol has also been used to determine change in tophus burden. The reported approach (Computer-Assisted Photographic Evaluation in Rheumatology, CAPER) has specified up to 5 measurable tophi (10). Using electronic calipers, the longest axis is measured together with the orthogonal axis to produce a measurement of area. Measurable tophi are defined as ≥5 mm in their longest dimension and to have distinguishable borders. In addition, up to 2 nonmeasurable tophi could be assessed qualitatively if they were ≥10 mm in their largest dimension (Table 1). The reported scoring system was the categories at the patient level (complete response, partial response, stable disease, and progressive disease) based on the definitions in Table 1 (10).

Number of items. Not applicable.

Response options/scale. Not applicable.

Recall period for items. Not applicable.

Endorsements. These techniques have not been unequivocally endorsed by any group. However, during OMERACT 10, 56 of 68 (82%) of the nonundecided participants agreed that the Vernier calipers method met the OMERACT filter for truth, discrimination, and feasibility. There were 37 additional participants who voted "Don't know" (11).

Examples of use. Enumeration of tophi has been used in randomized clinical trials of febuxostat and allopurinol (12,13), which showed that the number of tophi decreased after 40 months of effective urate-lowering therapy. In these trials, the tape measure method of a sentinel tophus has also shown change after prolonged normalization of serum urate levels. The Vernier calipers method has been used in a study that compared tophus size obtained from computed tomography. There was strong correlation between the 2 measurement techniques (14). In a longitudinal observational study, the Vernier calipers method was used to show that the velocity of tophus regression correlated strongly with the degree of urate lowering (15).

The digital photography method has been used in 2 replicate trials of pegloticase where it was shown that tophi regressed significantly after 12 weeks of therapy (10).

Practical Application

How to obtain. The methods of Vernier calipers, enumeration, and tape measurement are available for public use and are described clearly in a recent review (16). The digital photography method was developed by Savient Pharmaceuticals and RadPharm. Details of how to use this approach are available from Steve Hamburger (E-mail: shamburger@savientpharma.com).

Method of administration. These techniques are observer administered by direct examination.

Scoring. This is explained in the Content section and Table 2.

Score interpretation. Higher scores indicate greater tophus burden.

Respondent burden. Not applicable.

Administrative burden. The enumeration method, Vernier calipers, and tape measurement method are rapid (5 minutes) and require minimal or no equipment. The digital photography method requires a high up-front payment for the initial equipment (approximately $500) then low repeat costs. A training manual and video are available.

Image acquisition takes 5 to 7 minutes and image analysis takes up to 35 minutes depending on the number of tophi.

Translations/adaptations. Not applicable.

Psychometric Information

Method of development. These methods were developed as part of an effort to demonstrate changes in tophi in response to treatment.

Acceptability. Not reported.

Reliability. The reliability of the enumeration method and the digital photography method has not been reported. The intraobserver reliability for the tape measure method was ICC 0.92 (95% confidence interval [95% CI] 0.88–0.94), mean ± SD −0.2 ± 835 mm^2. The interobserver reliability (site 1) was ICC 0.92 (95% CI 0.86–0.96), mean ± SD −150 ± 982 mm^2; (site 2) ICC 0.85 (95% CI 0.75–0.91), mean ± SD 7 ± 925 mm^2 (17). The intraobserver reliability for the Vernier calipers method was ICC 1.0 (95% CI 0.99–1.0), mean ± SD −0.72 ± 2.42 mm. The interobserver reliability was ICC 0.99 (95% CI 0.97–0.99), mean ± SD 0.45 ± −2.3 mm (14).

Validity. All methods have good face validity. Only the Vernier calipers method has been compared with other methods of tophus measurement to establish construct validity (14). The study showed that there was a very high correlation ($r = 0.91$, $P < 0.0001$) between measures obtained by computed tomography and the Vernier calipers method. There was no difference in the coefficient of variation in measures obtained by either method. Only subcutaneous tophi were assessed by these methods and microscopic confirmation that the measured nodules were in fact tophi has not been obtained.

Ability to detect change. Each method has been shown to change in response to effective urate-lowering therapy. For the enumeration method, the mean percentage reduction in the total number of tophi was 58.5% after up to 3 years of treatment with an effect size of 0.47. Furthermore, a small but significant difference in the mean percent decrease in the number of tophi was observed with febuxostat 120 mg (−1.2) compared to placebo (−0.3) at 28 weeks ($P < 0.05$) (18). Using the tape measure method, tophus size was reduced by 59% and the effect size was 0.48 (13). A between-group difference in the change in tophus size has not been demonstrated with this method.

In a longitudinal observational study over 5 years, the Vernier calipers method showed that the velocity of tophus regression ranged from 0.57 to 1.53 mm/month and an effect size of 1.83 was observed (15). In this same study the velocity of tophus regression was greater in patients treated with benzbromarone. In the clinical trials that employed the digital photography method, 40% of patients experienced complete resolution of tophi and higher rates of complete resolution were observed in patients treated with pegloticase compared to placebo (7%; $P = 0.002$) (11).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Physical measurements of tophi are generally quick and easy to perform and are able to demonstrate change over time in people treated with effective urate-lowering therapy.

Caveats and cautions. Only palpable, subcutaneous tophi are observable by physical measurement. However, intrasynovial or periarticular tophi are more likely to be responsible for joint damage in gout and it is not currently clear that change in subcutaneous tophi will mirror change in unobservable tophi, although this seems likely. Another important unresolved issue is whether there are particular sites at which sentinel tophus assessment is more or less reliable. In addition, the minimally important change in tophus size or number has not been determined.

Clinical usability. The Vernier calipers and tape measurement methods are easily accommodated in the clinical setting. The observer reliability of these measures is sufficiently high to justify their use in the clinical setting. The enumeration method and digital photography cannot be recommended for routine clinical use at present, mainly because their observer and retest reliability are not published.

Research usability. The method with the most complete information regarding psychometric properties is the Vernier calipers method. This method received sufficient endorsement at OMERACT 10 to recommend this method for clinical research, although it has not yet been employed in the context of a randomized clinical trial.

AUTHOR CONTRIBUTIONS

Dr. Taylor drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Schumacher HR Jr, Edwards LN, Perez-Ruiz F, Becker M, Chen LX, Furst DE, et al. Outcome measures for acute and chronic gout. J Rheumatol 2005;32:2452–5.
2. Colwell HH, Hunt BJ, Pasta DJ, Palo WA, Mathias SD, Joseph-Ridge N, et al. Gout Assessment Questionnaire: initial results of reliability, validity and responsiveness. Int J Clin Pract 2006;60:1210–7.
3. Hirsch JD, Lee SJ, Terkeltaub R, Khanna D, Singh J, Sarkin A, et al. Evaluation of an instrument assessing influence of gout on health-related quality of life. J Rheumatol 2008;35:2406–14.
4. Lee SJ, Hirsch JD, Terkeltaub R, Khanna D, Singh JA, Sarkin A, et al. Perceptions of disease and health-related quality of life among patients with gout. Rheumatology (Oxford) 2009;48:582–6.
5. Singh JA, Sarkin A, Shieh M, Khanna D, Terkeltaub R, Lee SJ, et al. Health care utilization in patients with gout. Semin Arthritis Rheum 2010. E-pub ahead of print.
6. Sarkin AJ, Levack AE, Shieh MM, Kavanaugh AF, Khanna D, Singh JA, et al. Predictors of doctor-rated and patient-rated gout severity: gout impact scales improve assessment. J Eval Clin Pract 2010;16:1244–7.
7. Burnand B, Kernan WN, Feinstein AR. Indexes and boundaries for "quantitative significance" in statistical decisions. J Clin Epidemiol 1990;43:1273–84.
8. Singh JA, Taylor WJ, Simon LS, Khanna P, Stamp L, McQueen FM, et al. Patient reported outcomes in chronic gout: a report from OMERACT10. J Rheumatol 2011;38:1452–7.
9. Taylor WJ, Schumacher HJ, Baraf HS, Chapman P, Stamp L, Doherty M, et al. A modified delphi exercise to determine the extent of consensus with OMERACT outcome domains for studies of acute and chronic gout. Ann Rheum Dis 2008;67:888–91.
10. Maroli AN, Waltrip R, Alton M, Baraf HS, Huang B, Rehrig C, et al. First application of computer-assisted analysis of digital photographs for assessing tophus response: phase 3 studies of pegloticase in treatment failure gout [abstract]. Arthritis Rheum 2009;60:S416.
11. Dalbeth N, McQueen FM, Singh JA, MacDonald PA, Edwards NL, Schumacher HR Jr, et al. Tophus measurement as an outcome measure for clinical trials of chronic gout: progress and research priorities. J Rheumatol 2011;38:1458–61.

12. Becker MA, Schumacher HR Jr, Wortmann RL, MacDonald PA, Eustace D, Palo WA, et al. Febuxostat compared with allopurinol in patients with hyperuricemia and gout. N Eng J Med 2005;353:2450−61.
13. Becker MA, Schumacher HR, Macdonald PA, Lloyd E, Lademacher C. Clinical efficacy and safety of successful longterm urate lowering with febuxostat or allopurinol in subjects with gout. J Rheumatol 2009;36: 1273−82.
14. Dalbeth N, Clark B, Gregory K, Gamble GD, Doyle A, McQueen FM. Computed tomography measurement of tophus volume: comparison with physical measurement. Arthritis Rheum 2007;57:461–5.
15. Perez-Ruiz F, Calabozo M, Pijoan JI, Herrero-Beites AM, Ruibal A. Effect of urate-lowering therapy on the velocity of size reduction of tophi in chronic gout. Arthritis Rheum 2002;47:356−60.
16. Dalbeth N, Schauer C, McDonald P, Perez-Ruiz F, Schumacher HR, et al. Methods of tophus assessment in clinical trials of chronic gout: a systematic literature review and pictorial reference guide. Ann Rheum Dis 2011;70:597−604.
17. Schumacher HR Jr, Becker MA, Palo WA, Streit J, Macdonald PA, Joseph-Ridge N, et al. Tophaceous gout: quantitative evaluation by direct physical measurement. J Rheumatol 2005;32:2368−72.
18. Schumacher HR Jr, Becker MA, Wortmann RL, MacDonald PA, Hunt B, Streit J, et al. Effects of febuxostat versus allopurinol and placebo in reducing serum urate in subjects with hyperuricemia and gout: a 28-week, phase III, randomized, double-blind, parallel-group trial. Arthritis Rheum 2008;59:1540−8.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S297–S307
DOI 10.1002/acr.20638

Measures of Adult General Functional Status

SF-36 Physical Functioning Subscale (PF-10), Health Assessment Questionnaire (HAQ), Modified Health Assessment Questionnaire (MHAQ), Katz Index of Independence in Activities of Daily Living, Functional Independence Measure (FIM), and Osteoarthritis-Function-Computer Adaptive Test (OA-Function-CAT)

DANIEL K. WHITE, JESSICA C. WILSON, AND JULIE J. KEYSOR

INTRODUCTION

Self-reported measures to assess and quantify functional status are important tools for clinicians and investigators. These measures qualify limitation with different types of functional activities and quantify the extent of limitation. We are particularly interested in measures of general functional status. Although these "generic" measures of function were originally developed in other patient populations, they are relevant to the field of rheumatology. In particular, these instruments have been found to be valid and reliable measures of function, sensitive to changes in function, and have distinct thresholds for important change in people with rheumatologic disease.

Some notable studies have been added to the literature for general functional status measures in the last decade. Most of these additions are in the area of identifying thresholds for minimum clinically important difference, i.e., the smallest amount of change associated with a minimally important decline or improvement in function. To reflect changes in clinical practice over the last decade, we chose to review the Functional Independence Measure, which is a commonly used measure in practice to assess function. Also, Computer Adaptive Testing has developed over the past decade, which represents an innovative and exciting change to how self-reported tests of function are administered. Therefore, the purpose of this report is to provide an update to measures of general functional status commonly employed for people with rheumatologic diseases and provide a review of a Computer Adaptive Testing measure of functioning for people with osteoarthritis.

Dr. White's work was supported by an American College of Rheumatology Research and Education Foundation Rheumatology Investigator Award, the Boston Claude D. Pepper Older Americans Independence Center (P30-AG031679), and the Foundation for Physical Therapy.

Daniel K. White, PT, ScD, Jessica C. Wilson, PT, Julie J. Keysor, PT, PhD: Boston University, Boston, Massachusetts.

Address correspondence to Daniel White, PT, ScD, Clinical Epidemiology Research and Training Unit, 650 Albany Street, X200, Boston, MA 02118. E-mail: dwtbn@bu.edu.

Submitted for publication January 24, 2011; accepted in revised form May 10, 2011.

SF-36 PHYSICAL FUNCTIONING SUBSCALE (PF-10)

Description

Purpose. The PF-10 is a generic outcome measure designed to examine a person's perceived limitation with physical functioning (1) and is a subscale within the Medical Outcomes Study 36-item Short Form Health Survey (SF-36).

Content. Subjects are asked if their health limits physical activity, basic mobility, and basic activities of daily living.

Number of items in scale. There are 10 items.

Response options/scale. Responses are rated on a Likert scale. For the SF-36 versions 1.0 and 2.0, each item is rated on a 3-point scale (yes, limited a lot; yes, limited a little; and no, not limited at all). For the Patient-Reported Outcomes Measurement Information System (PROMIS) version, each item is rated on a 5-point scale (not at all, very little, somewhat, quite a lot, and cannot do).

Recall period for items. Respondents are asked to rate limitation at present for most questions and over the past 4 weeks for other questions.

Endorsements. None.

Examples of use. The PF-10 was developed as a generic health outcome instrument within the SF-36 for a wide variety of medical conditions in people ages 14–61. The PF-10 has been applied to older adult populations, as well as to people with rheumatoid arthritis (RA), back pain, osteoarthritis (OA), and gout (2).

Practical Application

How to obtain. The PF-10 instrument, scoring manual, and license are available from QualityMetric at www.qualitymetric.com. There is a charge at different rates for commercial and academic use. The PROMIS version of the PF-10 is available for viewing at http://www.nihpromis.org/default.aspx.

Method of administration. Interviewer (in person or by telephone) or self-administered.

Scoring. Answers to each question are summed to produce raw scores and then transformed to a 0–100 scale.

Score interpretation. Higher scores represent better health status. For the SF-36 version 2.0, the total PF-10 score is standardized to a mean of 50. Population norms are available for the US (3) and the UK (4,5). World data for cross-cultural comparisons are available as well (6).

Respondent burden. Less than 10 minutes is needed to complete the instrument. Questions are worded at a sixth- to ninth-grade level.

Administrative burden. Less than 10 minutes is necessary to administer the instrument and a few minutes are needed to score the results via computer. No training is required.

Translations/adaptations. There are 2 versions of the PF-10: the original SF-36 version 1.0 and the updated SF-36 version 2.0. Most recently, the PROMIS created a 10-item physical functioning scale that has 5 of the same questions as the PF-10 versions pertaining to vigorous activities and basic mobility, and 5 questions pertaining to basic and instrumental activities of daily living (7). However, this review will focus on the SF-36 versions 1.0 and 2.0. The SF-36 versions of the PF-10 are available in more than 50 different languages. More information on the availability of the PF-10 in other languages can be found from the International Quality of Life Assessment project at www.iqola.org.

Psychometric Information

Method of development. PF-10 questions were selected to assess a variety of physical activities ranging from easy to strenuous. The questionnaire was first examined in a group of subjects participating in the Medical Outcomes Study (8).

Acceptability. Missing data are not common. The PF-10 was designed to have low ceiling and floor effects.

Reliability. High test–retest reliability has been found in people with RA (intraclass correlation coefficient [ICC] 0.93) (9) and low back pain (ICC 0.83–0.91) (10). High internal consistency has also been reported for older adults (Cronbach's α = 0.82) (11) and people with gout (Cronbach's α = >0.93) (12).

Validity. *Criterion validity.* The PF-10 has been found to be associated with both generic and disease-specific measures of functional outcome in a variety of rheumatologic patient populations. For subjects with hip or knee OA, Salaffi and colleagues reported a high correlation between the PF-10 and the Western Ontario and McMaster Universities Arthritis Index physical function subscale (r = −0.65) (13). Similarly, a moderate correlation between the PF-10 and the Timed Up and Go Test was reported in subjects following total hip or knee replacement (r = −0.34) (14). For people from Norway with RA, the PF-10 has been found to have strong correlations with the Modified Health Assessment Questionnaire (r = −0.69) and the Arthritis Impact Measurement Scale physical domain (r = −0.73) (15). Lastly, the PF-10 has been shown to be highly correlated with the Late Life Function and Disability Index in older adults (r = 0.74−0.88) (2).

Construct validity. The PF-10 has been found to measure a single or unidimensional index in subjects with chronic medical and psychiatric conditions from the US (16), and in people with psoriatic arthritis (17). The PF-10 was also found to measure a unidimensional index among subjects from the general population from 7 countries, including Denmark, Germany, Italy, the US, Sweden, The Netherlands, and the UK (18).

Ability to detect change. The PF-10 has been found to be a sensitive and responsive instrument to change in subjects with RA (9,19), spine pathology (10,20,21), and chronic medical and psychiatric diseases (22). In particular, the PF-10 was able to discriminate between groups of people with RA at different levels of improvement measured by the American College of Rheumatology criteria following a drug trial (19). Similarly, the PF-10 was sensitive to change in people with spine pathology undergoing physical therapy (10). Lastly, using data from subjects with chronic disease within the Medical Outcomes Study, McHorney and colleagues reported that the PF-10 had similar sensitivity to change regardless whether scores were Rasch-transformed or not (22).

The minimum clinically important difference (MCID) for the PF-10 has been examined in subjects with spine pathology, specifically intervertebral disc herniation. In this patient population, the MCID is reported as ranging between 5 and 30 for the PF-10 (20).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The PF-10 is an important instrument of general physical function relevant to the rheumatology community. It evaluates limitations in function common in people with rheumatology-related disease and can be used to evaluate changes following intervention, especially in people with RA and spine pathology.

Caveats and cautions. The psychometrics for the PF-10 have not been consistently investigated across all rheumatologic conditions. For instance, more work is needed to establish MCID thresholds for the PF-10 in people with RA.

Clinical usability. The psychometric evaluation of the PF-10 does support interpretation of scores for individual patients and can be employed in the clinic given the short administration time.

Research usability. The psychometric evaluation of the PF-10 does support use in intervention studies and observational studies.

HEALTH ASSESSMENT QUESTIONNAIRE (HAQ)

Description

Purpose. This review focuses on the HAQ disability index with an emphasis on the use of the HAQ as a measure of general function (23). The HAQ measures difficulty in performing activities of daily living. It is the most widely used functional measure in rheumatology. The HAQ was specifically developed for use among adults with arthritis, but it has since been used in a wide range of populations (24).

Content. Questions assessing difficulty over the past week in 20 specific functions that are grouped into 8 categories: dressing and grooming, arising, eating, walking, personal hygiene, reaching, gripping, and other activities.

Number of items. There are 20 items covering 8 subscales: dressing and grooming (2 items: dress yourself, including tying shoelaces and fastening buttons, and shampoo your hair); arising (2 items: stand up straight from an armless straight chair, get in and out of bed); eating (3 items: cut your meat, lift a full cup or glass to your mouth, open a new milk carton); walking (2 items: walk outdoors on flat ground, climb up 5 steps); personal hygiene (3 items: wash and dry your entire body, take a tub bath, get on and off the toilet); reaching (2 items: reach and get down a 5-pound object from just above your head, bend down to pick up clothing from the floor); gripping (3 items: open car doors, open jars that have been previously opened, turn faucets on and off); and other activities (3 items: run errands and shop, get in and out of a car, do chores such as vacuuming or yard work). In addition, the use of personal assistance, assistive aids, or devices is measured.

Response options/scale. Each item is rated from 0–3, where 0 = no difficulty, 1 = some difficulty, 2 = much difficulty, and 3 = unable to do. The highest score within a category is used as the category score. Dependence on physical assistance or equipment raises the category score to 2. The HAQ score is calculated as the mean of the 8 category scores. Scores range from 0–3 in increments of 0.125. The overall score is not calculated if fewer than 6 category scores are completed.

Recall period for items. The past week.

Endorsements. None.

Examples of use. The HAQ was developed for individuals with rheumatoid arthritis (RA) and osteoarthritis (OA).

Practical Application

How to obtain. The English version of the HAQ and the Patient-Reported Outcomes Measurement Information System (PROMIS) HAQ and scoring directions are provided free of charge at http://aramis.stanford.edu/.

Method of administration. Interviewer (in person or by telephone) or self-administered (paper or electronic touch-screen version). The touch-screen version is self-explanatory and accessible for people with reduced motor function (25).

Scoring. The HAQ is hand scored. Alternate methods of scoring have been developed (for example, scoring without taking use of assistance or aids into account [26] or using the mean category score instead of the highest score [27]), but these scoring methods have not gained wide use. Wolfe suggests that even if alternative scoring methods are used, the traditional score should also be calculated in order to be compare with published data (28).

Score interpretation. Higher scores reflect more activity limitation. The overall estimated normal HAQ score was 0.25, with an average of 0.18 for males and 0.28 for females within a general population sample of 1,530 of people age ≥30 years in Central Finland (29). Approximately one-third of the respondents reported some sort of disability (HAQ >0). The prevalence of rates of disability increase exponentially after age 50 years (29).

Respondent burden. Less than 10 minutes are needed to complete the HAQ. Questions are worded at a sixth- to ninth-grade level.

Administrative burden. Less than 10 minutes are needed to administer the HAQ, and less than 2 minutes are needed to score the HAQ. No training is necessary.

Translations/adaptations. Many adaptations and/or translations are available, including English (US, Canada, Australia), Belgian Flemish and French, Canadian French, Chinese (Cantonese, Hong Kong), Danish, French, German, Spanish (US, Spain, many Central and South American countries), Swedish, and Turkish. For a complete listing, see Bruce and Fries (30). A revised version of the HAQ, the HAQ-II, has been developed and contains 10 items (31). A PROMIS HAQ has been developed, which contains the same 20 items as the original HAQ, but they were qualitatively improved to increase the clarity and psychometric properties of the measure (32).

Psychometric Information

Method of development. The HAQ was originally developed by using questions from a variety of instruments employed in the 1970s (23).

Acceptability. Missing data are not common. The HAQ has ceiling limitations, i.e., people with mild functional limitation can have normal HAQ scores.

Reliability. High test–retest reliability has been found in subjects with gout. Specifically, the test–retest reliability of the entire HAQ was intraclass correlation coefficient (ICC) 0.76, with individual subscales ranging from ICC 0.68 to ICC 0.80 (33). High correlations between interviewer versus self-administered forms of the instrument have been reported (range 0.60–0.88) (24), as well as between a touch-screen and paper version (ICC 0.99) (25).

Validity. For criterion validity, Daltroy et al (34) found a strong correlation (−0.72) between HAQ scores and a physical capacity measure in older adults.

For construct validity, HAQ scores are comparable across people with RA, OA, or gout using item response theory, which suggests the HAQ measures a single underlying construct of disability (35). Several studies have shown significant correlations of HAQ scores with other measures of function (e.g., Arthritis Impact Measurement Scale and Western Ontario and McMaster Universities Arthritis Index [WOMAC]) supporting the HAQ as a valid measure of general function (30,36–38).

Ability to detect change. The HAQ is a sensitive and responsive measure to changes in function in people with knee or hip OA. For people undergoing hip or knee joint replacement, the HAQ is responsive to functional change following surgery (39). Similarly, the HAQ has been found to be more sensitive to change over 3 years in people with hip or knee OA than the WOMAC (38). The HAQ has been found to have a ceiling effect, i.e., it does not discriminate well between people with low levels of disability (31,40). The minimum clinically important difference (MCID) for the HAQ has been examined in a variety of rheumatologic-related populations including RA, psoriatic arthritis, systemic lupus erythematosus, spondylarthropathies, and scleroderma. The range for MCID is −0.08 to −0.25 for improvement and 0.13 to 0.22 for decline (41–47). Several authors have commented that MCID values may depend on the severity of disability. Specifically, less change was needed to meet a meaningful threshold for improvement for people with low levels of disability compared with those with a high level of disability (41,43,46).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The HAQ measures important limitations in function relevant to many people with rheumatology-related disorders. Given that MCID values have been established for multiple rheumatologic populations, the HAQ is appropriate for evaluating interventions. It is notable that all studies investigating MCID reported a similar range of values, making the HAQ a useful measurement of function and change in function.

Caveats and cautions. The reliability, validity, and responsiveness of the HAQ requires more investigation. Specifically, the psychometrics of the HAQ need to be established across more rheumatologic patient populations. Clinical investigators should be aware that the HAQ does have floor effects, and may be less responsive to change among individuals with low levels of disability.

Clinical usability. The psychometric evaluation of the HAQ does support interpretation of scores for individual patients with moderate to severe disability and can be employed in the clinic given the short administration time.

Research usability. The psychometric evaluation of the HAQ does support use in intervention studies and observational studies, although sensitivity to change will likely be limited in people with mild disability.

MODIFIED HEALTH ASSESSMENT QUESTIONNAIRE (MHAQ)

Description

Purpose. The MHAQ is a modified version of the HAQ (48).

Content. The number of specific activities queried is reduced from 20 to 8 (1 item is used from each of the 8 categories covered in the HAQ). The MHAQ has 4 subscales that assess degree of difficulty, satisfaction with function, change in function over the past 6 months, and perceived need for help with each activity. The degree of difficulty subscale is the most commonly used.

Number of items. There are 8 items (dressing, arising, eating, walking, hygiene, reaching, gripping, and getting in and out of car) repeated in each of the 4 subscales.

Response options/scale. For the difficulty subscale ("Are you able to. . .?"), the scale is 0 = without any difficulty, 1 = with some difficulty, 2 = with much difficulty, and 3 = unable to do. Any positive response regarding help or assistive devices raises the score to 2. For satisfaction ("How satisfied are you with your ability to. . .?"), 0 = satisfied and 1 = dissatisfied. For change in difficulty ("Compared to 6 months ago, how difficult is it now [this week] to. . .?"), 0 = less difficult now, 1 = no change, and 2 = more difficult now. For need for help (Do you need help to. . .?"), 0 = do not need help and 1 = need help. Scale scores are the mean of the scores on the 8 items within the scale: difficulty 0–3, satisfaction 0–1, change in function 0–2, and need for help 0–1.

Recall period for items. Up to 6 months.

Endorsements. None.

Examples of use. People with rheumatic conditions (48).

Practical Application

How to obtain. Available in original reference (48).

Method of administration. Interviewer or self-administered.

Scoring. Arithmetic calculation by hand.

Score interpretation. Higher scores reflect poorer health.

Respondent burden. Less than 5 minutes are needed to complete the MHAQ. Questions are worded at a sixth- to ninth-grade level.

Administrative burden. Less than 5 minutes are needed to administer the MHAQ, and less than 2 minutes are needed to score the MHAQ. No training is necessary.

Translations/adaptations. Two subsequent versions of the HAQ have been developed, the Multidimensional Health Assessment Questionnaire (48), and the HAQ-II (31). Both instruments were developed to address ceiling problems associated with the MHAQ (40).

Psychometric Information

Method of development. Questions from the MHAQ are directly from the HAQ.

Acceptability. Missing data are not common. The MHAQ has floor limitations and ceiling limitations (48).

Reliability. The test–retest reliability for the difficulty scale over 1 month was reported as 0.91 (48).

Validity. For concurrent validity for the difficulty scale, the MHAQ is highly correlated with the overall score of the HAQ (0.88), the Arthritis Impact Measurement Scale physical component (0.80), and the Short Form 36 physical function scale (0.71) in people with rheumatoid arthritis (40). Blalock and colleagues also examined the equivalency of the HAQ with the MHAQ, and found that although the scores were highly correlated, the MHAQ scores were consistently and significantly lower (indicating better function) than the HAQ score (49). Uhlig and coauthors also found large numerical differences in scores, especially at higher disability levels (40). In every category, HAQ items chosen for the MHAQ had a lower mean than the MHAQ-excluded items (49). For construct validity for the difficulty scale, the MHAQ scores have been found to be associated with measures of physical performance (e.g., walk test, grip strength) (50). For construct validity for dissatisfaction with function scale, scores were incrementally greater (more dissatisfied) as difficulty in function increased (48).

Ability to detect change. For the difficulty scale, Blalock and colleagues suggest that the MHAQ is relatively insensitive to low levels of disability, and because of its restricted range and skewed distribution, should be used with caution when the intent is to assess functional change (49). Uhlig et al also reported a considerable ceiling effect for the MHAQ (40). Stucki et al (scores $<$0.3 [51]) and Wolfe (scores $\leq$1.0 [28]) also noted clustering of scores at the low end of the scale. Ziebland et al found that the MHAQ change in difficulty scale was more sensitive to changes in clinical variables (i.e., correlated more highly with variables such as grip strength, pain, morning stiffness, and erythrocyte sedimentation rate) than a pre-post difference in the traditional HAQ score (52).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Similar to the HAQ, the MHAQ measures important limitations in function relevant to many people with rheumatology-related disorders. Given its abbreviated form, the MHAQ should be considered when the full version of the HAQ cannot be implemented due to time constraints.

Caveats and cautions. The majority of psychometric analysis of the MHAQ has focused on the difficulty subscale, and has generally found that it appears to be less psychometrically sound than the HAQ. Blalock et al noted that scores on the MHAQ were consistently lower than those on the HAQ (49). Mean differences on the overall difficulty score were 0.67 lower using HAQ scores calculated with adjustment for help and/or assistive devices, and 0.52 lower using HAQ scores without such adjustments. The MHAQ does not make adjustments for use of help or assistive devices. Blalock also noted that while the HAQ scores were normally distributed across the scale's full possible range (0–3), MHAQ scores were not normally distributed and ranged only from 0–1.75. Similar findings were also noted by Stucki et al (51) and Wolfe (28). The MHAQ also has a considerable ceiling effect, which is greater than that of the HAQ (40). There are conflicting reports about correlations between MHAQ scores and clinical and laboratory variables. Wolfe concluded that the advantages in the length of the MHAQ over the HAQ were offset by loss of sensitivity and responsiveness to change (28).

Clinical usability. The psychometric evaluation of the MHAQ does support limited use in the clinic, however, floor and ceiling effects should be considered when interpreting scores.

Research usability. Given the MHAQ's limited ability to detect change, research use is not recommended.

KATZ INDEX OF INDEPENDENCE IN ACTIVITIES OF DAILY LIVING

Description

Purpose. To quantify independence in activities of daily living (ADL) across a wide range of patient populations (53).

Content. Basic ADL (bathing, dressing, toileting, transfers, continence, and feeding). Katz et al noted that the loss of functional skills occurs in a specific order, with the most complex lost first (54). The scoring method for this scale reflects this hierarchy of function.

Number of items. 6, 1 for each ADL.

Response options/scale. Each ADL is scored on a 3-point scale of independence. Items are ordered by difficulty. The scoring reflects this, although some variation in the hierarchy of difficulty is allowed. Katz reported that ADL functions of 86% of evaluated subjects were consistent with the hierarchy (54). Score range is A–G or 0–6.

Recall period for items. Immediate.

Endorsements. None.

Examples of use. The Katz Index of ADL has been used in older adults (55), people with stroke (56), and older adults with hip fracture (57).

Practical Application

How to obtain. Available from original reference (54) and at www.npcrc.org/resources/resources_show.htm?doc_id=376169.

Method of administration. Examiner-administered via observation of the patient.

Scoring. Independence in various combinations of ADL determines ordinal rank on the alpha scale, or the number of ADLs for which the individual is dependent for the numeric scale. Ratings are made are on an 8-level ordinal scale, where A = independence in feeding, continence, transferring, going to toilet, dressing, and bathing; B = independent in all but 1 of these functions; C = independent in all but bathing and 1 additional function; D = independent in all but bathing, dressing, and 1 additional function; E = independent in all but bathing, dressing, going to toilet, and 1 additional function; F = independent in all but bathing, dressing, going to toilet, transferring, and 1 additional function; G = dependent in all 6 functions; and other = dependent in at least 2 functions, but not classifiable as C, D, E, or F. Katz and Akpom later proposed a simplified scoring system in which individuals are scored 0–6, reflecting the number of ADLs in which they are dependent (58).

Score interpretation. Scores reflect the specific ADLs or number of dependent ADLs. Higher (alphabetically or numerically) scores reflect greater independence.

Respondent burden. Five minutes to complete. Instrument is performance based.

Administrative burden. Must observe the patient in each ADL to determine level of independence.

Translations/adaptations. The Katz Index of ADL has been adapted into several versions that are comparable to the original (59,60), while others have been modified

(61,62). The Katz Index of ADL has also been translated into Spanish (63).

Psychometric Information

Method of development. The Katz Index of ADL was developed from the observations of inpatients with hip fractures. Observations were made by physicians, nurses, and other health professionals (54).

Acceptability. The Katz Index of ADL measures only basic ADLs, and therefore has ceiling effects, i.e., the index cannot discriminate well among people with no and mild limitations.

Reliability. The interrater reliability is 0.95 or better after training (54,64). The coefficient of reproducibility (a measure of the internal consistency of an ordered measure) is 0.96–0.99 (65). In a study examining the reliability and validity of self-reported limitations in ADL among Turkish, Moroccan, and indigenous Dutch elderly in The Netherlands, Reijneveld et al reported that internal consistency reliabilities were good for all ethnic groups, being slightly higher for Turkish and Moroccan elderly people than for Dutch elderly (66).

Validity. Regarding construct validity, the Katz Index of ADL is associated with scores from the Barthel Index ($r = 0.78$ [67], $\kappa = 0.77$ [68]). The Spanish versions of the Katz Index of ADL are associated with mortality, institutionalization, and utilization of social health services (63). For predictive validity, the Katz Index of ADL is associated with mobility dysfunction (0.50) and house confinement (0.39) among older patients 2 years later (69). There is also a correlation between ADL dependency level and mortality among nursing home residents (64). Comparing patients at 1-month poststroke, those with grade A-B-C at admission were more likely to go home compared with those with a grade of D-E-F-G (56).

Ability to detect change. The scale had a significant floor effect, in that it is relatively insensitive to variations at low levels of disability (36). Scores on the Katz ADL scale are dependent on the physical environment, i.e., different scores may be obtained for individuals in different settings or with different environmental modifications (37).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The Katz Index of ADL measures important functional limitations, which can occur in rheumatologic patient populations.

Caveats and cautions. There has been little investigation of sensitivity and responsiveness of the Katz Index of ADL. Most problematic is potential for ceiling effects with people with mild limitations in ADLs. This could lead to the index not being responsive to changes in ADLs in people with low levels of disability.

Clinical usability. The psychometric evaluation provides some support for the clinical use of the Katz Index of ADL, however, more robust measures of ADL function, such as the Functional Independence Measure should be considered.

Research usability. Use of the Katz Index of ADL in research studies is not well supported.

FUNCTIONAL INDEPENDENCE MEASURE (FIM)

Description

Purpose. The FIM estimates the level of assistance needed for patients to complete basic activities of daily living (ADL) (70). The FIM was designed to be an assessment tool that could be implemented universally across all patient populations within an inpatient rehabilitation hospital environment (70).

Content. The FIM includes 18 basic ADLs, such as self-care, sphincter control, transfers, locomotion, communication, and social cognition. Clinicians score patients on a 7-point scale ranging from dependent to independent, which reflects the level of assistance needed to complete each ADL.

Number of items. The FIM items are organized into the motor and cognitive domains, which are further organized into 4 subscales for the motor domain and 2 subscales for the cognitive domain.

Response options/scale. A trained health professional rates a patient on a scale of 1–7, where 1 = total assistance (the patient provides <25% effort to complete each task), 2 = maximal assistance (25–49% effort), 3 = moderate assistance (50–74% effort), 4 = minimal assistance (>75% effort), 5 = supervision/set up (need for supervision but no physical contact), 6 = modified independence (use of a device or need for more than a reasonable time to complete each task), and 7 = complete independence (the patient completes each task in a timely and safe manner). Different health professionals can score sections specific to their discipline. For instance, a physical therapist can score the mobility-related items for a patient while an occupational therapist scores the ADL-related items. There are 2 gross score classifications: dependent (helper: scores 1–5) and independent (no helper: scores 6–7). The total FIM score is calculated by summing the score of each of the 18 items.

Recall period for items. Immediate.

Endorsements. The FIM is used to determine payment for inpatient acute rehabilitation services from the Centers for Medicare and Medicaid Services. In particular, the FIM is used to determine coverage for patients under Medicare Part A.

Examples of use. Individuals within the inpatient acute rehabilitation hospital setting.

Practical Application

How to obtain. The FIM System program is available at http://www.udsmr.org/. A sample of the FIM instrument can be found at http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=physmedrehab&part=A11332&rendertype=figure&id=A11340.

Method of administration. Observation by members of an interdisciplinary team.

Scoring. Specific scoring instructions apply to the FIM. Training manuals for scoring are available from the Centers from Medicare and Medicaid Services, http://www.cms.gov/InpatientRehabFacPPS/04_IRFPAI.asp.

Score interpretation. Scores range from 18–126. Higher scores represent more independence. A score of 18 represents complete dependence, while a score of 126 represents complete independence. The total FIM score is appropriate to report if the goal of assessment is to determine the overall burden of care (71). There are 2 domains of the FIM: motor and cognitive. The motor domain subscales include self-care (6 items: eating, grooming, bathing, dressing upper body, dressing lower body, and toileting); sphincter control (2 items: bladder management, bowel management); transfers (3 items: bed/chair/wheelchair, toilet, tub/shower); and locomotion (2 items: walk or wheelchair, stairs). The motor domain was developed from the Barthel Index (72). The cognitive domain subscales include communication (2 items: comprehension, expression) and social cognition (3 items: social interaction, problem solving, memory). The mean ± SD admission FIM total was 73.2 ± 12.9 and discharge FIM total was 101.7 ± 12.9 for patients with lower extremity joint replacement who were discharged from a rehabilitation program in 2007 (73).

Respondent burden. 30–45 minutes to perform all activities. Patients are asked to perform each functional task in order to generate a score, which may be difficult.

Administrative burden. 7 minutes to collect demographic data and 10 minutes to score. Formal training is needed to administer the FIM. A training examination is available at: http://www.udsmr.org/.

Translations/adaptations. The FIM has been translated into different languages including Italian and Turkish (74,75).

Psychometric Information

Method of development. The FIM was created to provide an improvement over the Barthel Index. It has been developed and tested mainly in people with neurologic pathology.

Acceptability. Missing data are not common. The instrument has some ceiling effects within each of the motor and cognitive domains.

Reliability. High reliability has been reported for the FIM. In a quantitative review of 11 studies, Ottenbacher et al reported high interrater and test–retest reliability for health professionals with a variety of educational backgrounds and levels of training (76). Based on 1,568 patients with a variety of medical diagnoses, the median interrater reliability was 0.95 and test–retest reliability was 0.95. Median reliability for the 6 subscales ranged from 0.78 (social cognition) to 0.95 (self-care), and the 18 individual items ranged from 0.61 (comprehension) to 0.90 (toilet transfer). Pollak and colleagues also found high test–retest reliability for the motor (intraclass correlation coefficient [ICC] 0.90) and cognitive domains (ICC 0.80) in a cohort of older adults age ≥80 years residing in a multilevel retirement community (77).

High internal consistency was found for the total FIM score (Cronbach's $\alpha = 0.88-0.97$) (71,78), the motor domain ($\alpha = 0.84-0.97$) (71,79), and the cognitive domain ($\alpha = 0.86-0.95$) (71) within a large sample of inpatients undergoing acute rehabilitation with various diagnoses. However, lower internal consistency was reported for the locomotion subscale ($\alpha = 0.68$), suggesting that the individual items (ambulation/wheelchair use and stair climbing) may be measuring a different latent construct of function (78). Internal consistency was also high for FIM scores obtained via interview ($\alpha = 0.94$) or observation ($\alpha = 0.90$) (80).

Validity. Regarding concurrent validity, FIM scores assigned by a single nonclinician interview and by observation by a team of health care professionals were similar (ICC 0.74 for admission FIM and ICC 0.76 for discharge FIM), which provides evidence that a multi-interviewer–administered FIM is a valid method for collecting data (81). For construct validity, the separation of the FIM into motor and cognitive domains has been found to be a valid method of measuring activity limitation (82–84). The items in each domain show a generally consistent pattern of difficulty rating across multiple medical diagnoses, with eating the least difficult motor item to achieve an independent rating, and stair climbing the most difficult (82–84). For cognitive items, expression is the least difficult and problem solving is the most difficult (82,84). FIM scores are correlated with age, comorbidity, and discharge destination (78), as well as other functional measures, such as the Barthel Index and the Functional Assessment Measure (79,80,85,86).

While little work has examined the predictive validity of the FIM within rheumatologic patient populations, several studies have examined this within stroke. Trends from these studies can be carefully considered for patients with rheumatologic conditions who are at an inpatient rehabilitation hospital. Admission FIM scores have been shown to predict length of stay and discharge FIM scores in a rehabilitation hospital following stroke (80,87–91). In particular, an increase in the admission score of the motor domain by 1 point is correlated with a 1.1-day decrease in average rehabilitation length of stay for patients with stroke (87). There is a strong association between total FIM scores and discharge destination, i.e., discharge home versus skilled nursing facility (90,92–95). A majority of patients with stroke with admission FIM scores >80 are discharged home, while less than half with admission FIM scores <40 are discharged home, regardless of age (94). Social support has been shown to be a decisive factor for discharge destination, especially for those requiring high levels of assistance (90,93).

Ability to detect change. The FIM, especially the motor FIM, is highly responsive in detecting changes in ADL performance (78,80,96), but the cognitive FIM has a poor responsiveness due to its significant ceiling effect seen across a wide variety of medical diagnoses (96–99). There is comparable responsiveness between the FIM and the Barthel Index (79,80,85,96,100). Beninato et al reported a minimum clinically important difference for the total FIM of 22, the motor FIM of 17, and the cognitive FIM of 3 in the stroke population when anchored to a physician's assessment of minimally clinically important change (101).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The FIM is a widely used tool in the rehabilitation setting across a broad range of medical diagnoses including rheumatologic diagnoses. The FIM is appropriate for evaluating interventions for people with severe functional limitation.

Caveats and cautions. The FIM is not intended for community-dwelling adults who are independent in most functional activities. Future work is needed to validate the predictive validity of the FIM within rheumatologic patient populations.

Clinical usability. The psychometric evaluation does support interpretation of scores for individuals with severe functional limitation. Clinical use is primarily done in an inpatient acute rehabilitation setting.

Research usability. The psychometric evaluation does support use of the FIM within intervention studies and observational studies.

OSTEOARTHRITIS-FUNCTION-COMPUTER ADAPTIVE TEST (OA-FUNCTION-CAT)

Description

Purpose. The OA-FUNCTION-CAT employs computer adaptive testing to estimate a respondent's level of functioning. It was developed as a disease-specific measure for people with hip or knee OA (102).

Content. The OA-FUNCTION-CAT utilizes an item bank of 125 functional activities specific to hip or knee OA.

Number of items in scale. The OA-FUNCTION-CAT selects 5, 10, or 15 items from the 125-item bank for administration.

Recall period for items. Over the past month on an average day.

Endorsements. None.

Examples of use. The OA-Function-CAT was developed in a hip and knee OA cohort of subjects (102).

Practical Application

How to obtain. Contact CREcare (http://www.crecare.com/home.html) regarding cost and availability of the instrument. The 125-item bank is available for no fee at http://www.biomedcentral.com/content/supplementary/ar2760-S1.doc.

Method of administration. A CAT tailors assessment to each individual by selecting and administering subsequent questions based on the individual's response to the previous question. The program begins by selecting a question from the middle of the continuum of the calibrated item bank. Based on how the respondent answers the question, the computer calculates an initial score and level of precision. The CAT will conclude the test based on predetermined stop rules based on level of precision and/or a maximum number of items that are to be used to estimate the score. After the first question is answered, the program decides if the stop rule has been met. If not, another question is selected from the item bank based on the answer given for the previous question. This process is repeated until the stop rule has been satisfied, and a final score is calculated. This approach allows for the selection of items that provide the most relevant information at the level of the individual's current score estimate, therefore eliminating irrelevant questions from being asked (102–104).

Scoring. Continuous scale. For the functional difficulty scale, items are reported in terms of amount of difficulty in performing each function (none, a little, or a lot). For the functional pain scale, items are reported in terms of pain severity in performing each function (none, mild or moderate, or severe). The computer automatically calculates an outcome score representing how much limitation the individual has within the spectrum of functional limitation. This score is based on the individual's response to each of the questions asked.

Interpretation of scores. Scores range from 0–100. Higher scores represent higher function and less pain. The score produced on the CAT can be compared to other OA-FUNCTION-CAT scores regardless of the specific questions that were asked to generate the score. The OA-FUNCTION-CAT calculates a functional outcome score that can be compared within and between respondents.

Respondent burden. 15 or fewer questions are asked (questions written on a sixth-grade level of comprehension).

Administrative burden. Minimal burden since the computer program calculates the score in real time, so the score is available immediately.

Translations/adaptations. None.

Training to interpret. Not reported.

Psychometric Information

Reliability. There is high level of accuracy between the 5-, 10-, and 15-item OA-FUNCTION-CATs and the full item bank (Pearson's r = 0.92, 0.96, and 0.97, respectively, for the functional difficulty subscale and 0.89, 0.95, and 0.97, respectively, for the functional pain subscale) among people with hip or knee OA. There is high conditional reliability, i.e., examinee level reliability (105), for both the functional difficulty and functional pain subscales (95% of the sample scores achieved reliability estimates >0.97 and >0.96, respectively) (102).

Validity. Regarding construct validity, both the functional difficulty and functional domain subgroups fit a unidimensional model. Both of the OA-FUNCTION-CAT subscales cover a broader estimated scoring range than the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), especially at the upper, i.e., higher functioning, end of the scale. The OA-FUNCTION-CAT had less of a ceiling effect than the WOMAC (0.6% of subjects were at the ceiling for the OA-FUNCTION-CAT functional pain subscale versus 6.4% for the WOMAC pain scale, and 0.6% of subjects were at the ceiling for the OA-FUNCTION-CAT functional difficulty subscale versus 3.0% for the WOMAC physical function scale). The OA-FUNCTION-CAT did not have a floor effect (102).

Ability to detect change. The 10-item OA-FUNCTION-CAT has a higher degree of precision than the WOMAC across the full range of scores for both subscales, especially at the upper end of the scale in the functional pain subscale within people with hip or knee OA (102).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The OA-FUNCTION-CAT is an innovative method of measuring patient reported outcomes relevant to people with rheumatologic related disorders. The OA-FUNCTION-CAT has improved psychometric properties and requires fewer questions compared with legacy measures. Specifically, CATs offer a highly reliable and precise method to quantify patient reported limitations along a broad continuum. In addition, CAT scores can be estimated after only a few questions are answered, which decreases overall time and cost of administration.

Caveats and cautions. Future work is needed to examine the test–retest reliability of the OA-FUNCTION-CAT; utilization of CAT methods for estimating patient-reported outcomes is likely to increase among clinicians and researchers.

Clinical usability. The psychometric evaluation of the OA-FUNCTION-CAT supports interpretation of scores to make decisions about individuals. Given the minimal burden on patients and clinicians, the OA-FUNCTION-CAT is very appropriate to use clinically.

Research usability. The OA-FUNCTION-CAT can be used within intervention trials and observational studies given the psychometrics of this instrument. Values representing meaningful change have yet to be established which may limit clinical and research application.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Ware JE Jr, Sherbourne CD. The MOS 36-item Short-Form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
2. Dubuc N, Haley S, Ni P, Kooyoomjian J, Jette A. Function and disability in late life: comparison of the Late-Life Function and Disability Instrument to the Short-Form-36 and the London Handicap Scale. Disabil Rehabil 2004;26:362–70.
3. Ware JE Jr, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Med Care 1995;33 Suppl: AS264–79.
4. Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age. BMJ 1993; 306:1437–40.
5. Bowling A, Bond M, Jenkinson C, Lamping DL. Short Form 36 (SF-36) Health Survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, the Health Survey for England and the Oxford Healthy Life Survey. J Public Health Med 1999;21:255–70.
6. Ware JE Jr, Gandek B, Kosinski M, Aaronson NK, Apolone G, Brazier J, et al. The equivalence of SF-36 summary health scores estimated using standard and country-specific algorithms in 10 countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1167–70.
7. Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther 2009;11:R191.
8. Stewart AL, Ron DH, Ware JE Jr. The MOS Short-Form General Health Survey: reliability and validity in a patient population. Med Care 1988;26:724–35.
9. Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36). Br J Rheumatol 1998;37:425–36.
10. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 2002;82:8–24.
11. Bohannon RW, DePasquale L. Physical Functioning Scale of the Short-Form (SF) 36: internal consistency and validity with older adults. J Geriatr Phys Ther 2010;33:16–8.
12. Ten Klooster PM, Oude Voshaar MA, Taal E, van de Laar MA. Comparison of measures of functional disability in patients with gout. Rheumatology (Oxford) 2011;50:709–13.
13. Salaffi F, Carotti M, Grassi W. Health-related quality of life in patients with hip or knee osteoarthritis: comparison of generic and disease-specific instruments. Clin Rheumatol 2005;24:29–37.
14. Gandhi R, Tsvetkov D, Davey JR, Syed KA, Mahomed NN. Relationship between self-reported and performance-based tests in a hip and knee joint replacement population. Clin Rheumatol 2009;28:253–7.
15. Kvien TK, Kaasa S, Smedstad LM. Performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis II: a comparison of the SF-36 with disease-specific measures. J Clin Epidemiol 1998;51:1077–86.
16. Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10) I: unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol 1994;47:671–84.
17. Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. Arthritis Rheum 2007;57:723–9.
18. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1203–14.
19. Martin M, Kosinski M, Bjorner JB, Ware JE Jr, Maclean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res 2007;16:647–60.
20. Spratt KF. Patient-level minimal clinically important difference based on clinical judgment and minimally detectable measurement difference: a rationale for the SF-36 physical function scale in the SPORT intervertebral disc herniation cohort. Spine (Phila Pa 1976) 2009;34:1722–31.
21. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine (Phila Pa 1976) 1995;20:1899–908.
22. McHorney CA, Haley SM, Ware JE Jr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10) II: comparison of relative precision using Likert and Rasch scoring methods. J Clin Epidemiol 1997; 50:451–61.
23. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. 1982; 9:789–93.
24. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
25. Schefte DB, Hetland ML. An open-source, self-explanatory touch screen in routine care: validity of filling in the Bath measures on Ankylosing Spondylitis Disease Activity Index, Function Index, the Health Assessment Questionnaire and Visual Analogue Scales in comparison with paper versions. Rheumatology (Oxford) 2010;49:99–104.
26. Van der Heide A, Jacobs JW, van Albada-Kuipers GA, Kraaimaat FW, Geenen R, Bijlsma JW. Self report functional disability scores and the use of devices: two distinct aspects of physical function in rheumatoid arthritis. Ann Rheum Dis 1993;52:497–502.
27. Tomlin GS, Holm MB, Rogers JC, Kwoh CK. Comparison of standard and alternative health assessment questionnaire scoring procedures for documenting functional outcomes in patients with rheumatoid arthritis. J Rheumatol 1996;23:1524–30.
28. Wolfe F. Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. J Rheumatol 2001;28:982–9.
29. Krishnan E, Sokka T, Hakkinen A, Hubert H, Hannonen P. Normative values for the Health Assessment Questionnaire Disability Index:

benchmarking disability in the general population. Arthritis Rheum 2004;50:953–60.

30. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.
31. Wolfe F, Michaud K, Pincus T. Development and validation of the Health Assessment Questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum 2004;50:3296–305.
32. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061–6.
33. Alvarez-Hernandez E, Pelaez-Ballestas I, Vazquez-Mellado J, Teran-Estrada L, Bernard-Medina AG, Espinoza J, et al. Validation of the Health Assessment Questionnaire Disability Index in patients with gout. Arthritis Rheum 2008;59:665–9.
34. Daltroy LH, Larson MG, Eaton HM, Phillips CB, Liang MH. Discrepancies between self-reported and observed physical function in the elderly: the influence of response shift and other factors. Soc Sci Med 1999;48:1549–61.
35. Van Groen MM, ten Klooster PM, Taal E, van de Laar MA, Glas CA. Application of the Health Assessment Questionnaire disability index to various rheumatic diseases. Qual Life Res 2010;19:1255–63.
36. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 2nd ed. New York: Oxford University Press; 1996.
37. Spilker B. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven; 1996.
38. Bruce B, Fries J. Longitudinal comparison of the Health Assessment Questionnaire (HAQ) and the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). Arthritis Rheum 2004;51: 730–7.
39. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542–7.
40. Uhlig T, Haavardsholm EA, Kvien TK. Comparison of the Health Assessment Questionnaire (HAQ) and the modified HAQ (MHAQ) in patients with rheumatoid arthritis. Rheumatology (Oxford) 2006;45: 454–8.
41. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements: an illustration in rheumatology. Arch Intern Med 1993;153:1337–42.
42. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE Jr. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum 2000;43:1478–87.
43. Pope JE, Khanna D, Norrie D, Ouimet JM. The minimally important difference for the health assessment questionnaire in rheumatoid arthritis clinical practice is smaller than in randomized controlled trials. J Rheumatol 2009;36:254–9.
44. Kwok T, Pope JE. Minimally important difference for patient-reported outcomes in psoriatic arthritis: Health Assessment Questionnaire and pain, fatigue, and global visual analog scales. J Rheumatol 2010;37: 1024–8.
45. Colangelo KJ, Pope JE, Peschken C. The minimally important difference for patient reported outcomes in systemic lupus erythematosus including the HAQ-DI, pain, fatigue, and SF-36. J Rheumatol 2009;36: 2231–7.
46. Wheaton L, Pope J. The minimally important difference for patient-reported outcomes in spondyloarthropathies including pain, fatigue, sleep, and Health Assessment Questionnaire. J Rheumatol 2010;37: 816–22.
47. Sekhon S, Pope J, Baron M. The minimally important difference in clinical practice for patient-centered outcomes including health assessment questionnaire, fatigue, pain, sleep, global visual analog scale, and SF-36 in scleroderma. J Rheumatol 2010;37:591–8.
48. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
49. Blalock SJ, Sauter SV, Devellis RF. The modified Health Assessment Questionnaire difficulty scale: a health status measure revisited. Arthritis Care Res 1990;3:182–8.
50. Arvidson NG, Larsson A, Larsen A. Simple function tests, but not the modified HAQ, correlate with radiological joint damage in rheumatoid arthritis. Scand J Rheumatol 2002;31:146–50.
51. Stucki G, Stucki S, Bruhlmann P, Michel BA. Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. Ann Rheum Dis 1995;54: 461–5.
52. Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ. Ann Rheum Dis 1992;51:1202–5.
53. Katz P, for the Association of Rheumatology Health Professionals Outcomes Measures Task Force. Measures of adult general functional status. Arthritis Rheum 2003;49:S15–27.
54. Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. The Index of ADL: a standardized measure of biological and psychosocial function. JAMA 1963;185:914–9.
55. Asberg KH, Sonn U. The cumulative structure of personal and instrumental ADL: a study of elderly people in a health service district. Scand J Rehabil Med 1989;21:171–7.
56. Asberg KH, Nydevik I. Early prognosis of stroke outcome by means of Katz Index of activities of daily living. Scand J Rehabil Med 1991;23: 187–91.
57. Beloosesky Y, Grinblat J, Epelboym B, Weiss A, Grosman B, Hendel D. Functional gain of hip fracture patients in different cognitive and functional groups. Clin Rehabil 2002;16:321–8.
58. Katz S, Akpom CA. A measure of primary sociobiological functions. Int J Health Serv 1976;6:493–508.
59. Spector WD, Katz S, Murphy JB, Fulton JP. The hierarchical relationship between activities of daily living and instrumental activities of daily living. J Chronic Dis 1987;40:481–9.
60. Rodgers W, Miller B. A comparative analysis of ADL questions in surveys of older people. J Gerontol B Psychol Sci Soc Sci 1997;52 Spec No:21–36.
61. Reuben DB, Valle LA, Hays RD, Siu AL. Measuring physical function in community-dwelling older persons: a comparison of self-administered, interviewer-administered, and performance-based measures. J Am Geriatr Soc 1995;43:17–23.
62. LaPlante MP. The classic measure of disability in activities of daily living is biased by age but an expanded IADL/ADL measure is not. J Gerontol B Psychol Sci Soc Sci 2010;65:720–32.
63. Cabanero-Martinez MJ, Cabrero-Garcia J, Richart-Martinez M, Munoz-Mendoza CL. The Spanish versions of the Barthel index (BI) and the Katz index (KI) of activities of daily living (ADL): a structured review. Arch Gerontol Geriatr 2009;49:e77–84.
64. Spector WD, Takada HA. Characteristics of nursing homes that affect resident outcomes. J Aging Health 1991;3:427–54.
65. Brorsson B, Asberg KH. Katz index of independence in ADL: reliability and validity in short-term care. Scand J Rehabil Med 1984;16:125–32.
66. Reijneveld SA, Spijker J, Dijkshoorn H. Katz' ADL index assessed functional performance of Turkish, Moroccan, and Dutch elderly. J Clin Epidemiol 2007;60:382–8.
67. Rockwood K, Stolee P, Fox RA. Use of goal attainment scaling in measuring clinically important change in the frail elderly. J Clin Epidemiol 1993;46:1113–8.
68. Gresham GE, Phillips TF, Labi ML. ADL status in stroke: relative merits of three standard indexes. Arch Phys Med Rehabil 1980;61: 355–8.
69. Katz S, Downs TD, Cash HR, Grotz RC. Progress in development of the index of ADL. Gerontologist 1970;10:20–30.
70. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. Adv Clin Rehabil 1987;1:6–18.
71. Stineman MG, Shea JA, Jette A, Tassoni CJ, Ottenbacher KJ, Fiedler R, et al. The Functional Independence Measure: tests of scaling assumptions, structure, and reliability across 20 diverse impairment categories. Arch Phys Med Rehabil 1996;77:1101–8.
72. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. Md State Med J 1965;14:61–5.
73. Granger CV, Markello SJ, Graham JE, Deutsch A, Reistetter TA, Ottenbacher KJ. The uniform data system for medical rehabilitation: report of patients with lower limb joint replacement discharged from rehabilitation programs in 2000-2007. Am J Phys Med Rehabil 2010;89: 781–94.
74. Invernizzi M, Carda S, Milani P, Mattana F, Fletzer D, Iolascon G, et al. Development and validation of the Italian version of the Spinal Cord Independence Measure III. Disabil Rehabil 2010;32:1194–203.
75. Kucukdeveci AA, Yavuzer G, Elhan AH, Sonel B, Tennant A. Adaptation of the Functional Independence Measure for use in Turkey. Clin Rehabil 2001;15:311–9.
76. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. Arch Phys Med Rehabil 1996;77:1226–32.
77. Pollak N, Rheault W, Stoecker JL. Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. Arch Phys Med Rehabil 1996;77:1056–61.
78. Dodds TA, Martin DP, Stolov WC, Deyo RA. A validation of the functional independence measurement and its performance among rehabilitation inpatients. Arch Phys Med Rehabil 1993;74:531–6.
79. Hsueh IP, Lin JH, Jeng JS, Hsieh CL. Comparison of the psychometric characteristics of the functional independence measure, 5 item Barthel index, and 10 item Barthel index in patients with stroke. J Neurol Neurosurg Psychiatry 2002;73:188–90.
80. Sadaria KS, Bohannon RW, Lee N, Maljanian R. Ratings of physical

function obtained by interview are legitimate for patients hospitalized after stroke. J Stroke Cerebrovasc Dis 2001;10:79–84.

81. Young Y, Fan MY, Hebel JR, Boult C. Concurrent validity of administering the functional independence measure (FIM) instrument by interview. Am J Phys Med Rehabil 2009;88:766–70.
82. Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger C. Relationships between impairment and physical disability as measured by the functional independence measure. Arch Phys Med Rehabil 1993;74:566–73.
83. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil 1994;75:127–32.
84. Granger CV, Hamilton BB, Linacre JM, Heinemann AW, Wright BD. Performance profiles of the functional independence measure. Am J Phys Med Rehabil 1993;72:84–9.
85. Hobart JC, Lamping DL, Freeman JA, Langdon DW, McLellan DL, Greenwood RJ, et al. Evidence-based measurement: which disability scale for neurologic rehabilitation? Neurology 2001;57:639–44.
86. Gosman-Hedstrom G, Svensson E. Parallel reliability of the functional independence measure and the Barthel ADL index. Disabil Rehabil 2000;22:702–15.
87. Tan WS, Heng BH, Chua KS, Chan KF. Factors predicting inpatient rehabilitation length of stay of acute stroke patients in Singapore. Arch Phys Med Rehabil 2009;90:1202–7.
88. Inouye M, Kishi K, Ikeda Y, Takada M, Katoh J, Iwahashi M, et al. Prediction of functional outcome after stroke rehabilitation. Am J Phys Med Rehabil 2000;79:513–8.
89. Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger C. Prediction of rehabilitation outcomes with disability measures. Arch Phys Med Rehabil 1994;75:133–43.
90. Koyama T, Sako Y, Konta M, Domen K. Poststroke discharge destination: functional independence and sociodemographic factors in urban Japan. J Stroke Cerebrovasc Dis 2011;20:202–7.
91. Ng YS, Jung H, Tay SS, Bok CW, Chiong Y, Lim PA. Results from a prospective acute inpatient rehabilitation database: clinical characteristics and functional outcomes using the Functional Independence Measure. Ann Acad Med Singapore 2007;36:3–10.
92. Black TM, Soltis T, Bartlett C. Using the Functional Independence Measure instrument to predict stroke rehabilitation outcomes. Rehabil Nurs 1999;24:109–14, 121.
93. Lutz BJ. Determinants of discharge destination for stroke patients. Rehabil Nurs 2004;29:154–63.
94. Alexander MP. Stroke rehabilitation outcome: a potential use of predictive variables to establish levels of care. Stroke 1994;25:128–34.
95. Gulati A, Yeo CJ, Cooney AD, McLean AN, Fraser MH, Allan DB. Functional outcome and discharge destination in elderly patients with spinal cord injuries. Spinal Cord 2011;49:215–8.
96. Van der Putten JJ, Hobart JC, Freeman JA, Thompson AJ. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel index and the Functional Independence Measure. J Neurol Neurosurg Psychiatry 1999;66:480–4.
97. Davidoff GN, Roth EJ, Haughton JS, Ardner MS. Cognitive dysfunction in spinal cord injury patients: sensitivity of the Functional Independence Measure subscales vs neuropsychologic assessment. Arch Phys Med Rehabil 1990;71:326–9.
98. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the Functional Independence Measure in traumatic spinal cord injury. Arch Phys Med Rehabil 1999;80:1471–6.
99. Kohler F, Dickson H, Redmond H, Estell J, Connolly C. Agreement of functional independence measure item scores in patients transferred from one rehabilitation setting to another. Eur J Phys Rehabil Med 2009;45:479–85.
100. Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke: the impact of using different methods for measuring responsiveness. J Clin Epidemiol 2002;55:922–8.
101. Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. Arch Phys Med Rehabil 2006;87:32–9.
102. Jette AM, McDonough CM, Ni P, Haley SM, Hambleton RK, Olarsch S, et al. A functional difficulty and functional pain instrument for hip and knee osteoarthritis. Arthritis Res Ther 2009;11:R107.
103. Jette AM, Haley SM. Contemporary measurement techniques for rehabilitation outcomes assessment. J Rehabil Med 2005;37:339–45.
104. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res 2007;16 Suppl 1:133–41.
105. Raju NS, Price LR, Oshima TC, Nering ML. Standardized conditional SEM: a case for conditional reliability. Appl Psychol Meas 2007;31: 169–80.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S325–S336
DOI 10.1002/acr.20641

Measures of Social Function and Participation in Musculoskeletal Populations

Impact on Participation and Autonomy (IPA), Keele Assessment of Participation (KAP), Participation Measure for Post-Acute Care (PM-PAC), Participation Objective, Participation Subjective (POPS), Rating of Perceived Participation (ROPP), and The Participation Scale

ROSS WILKIE, JOANNE L. JORDAN, SARA MULLER, ELAINE NICHOLLS, EMMA L. HEALEY, AND DANIELLE A. VAN DER WINDT

Maintaining participation and social function is important to individuals with musculoskeletal conditions (1). Although there is no single clearly specified or widely agreed concept of participation, it is generally understood to be the domain of functioning beyond impairments and the performance of basic tasks (e.g., sit to stand) and refers to the context in which people live (2). The publication of the World Health Organization's International Classification of Functioning in 2001 has done much to raise awareness of the need to measure social function and participation, and their definition can be used as a starting point; participation, the concept proposed to capture social function, refers to the experience in life situations where an individual interacts with their environment (e.g., physical environment and other people) (3). Participation has evolved from a number of concepts, including handicap, instrumental activities of daily living, and social role participation, and is a broad term that covers an individual's experience in an infinite number of life activities and social roles, for example, work, looking after others, leisure activities, volunteering, and being involved in the community.

Participation and social function have not been measured routinely in musculoskeletal populations in clinical practice or research studies (4). Measurement of the consequences and management of musculoskeletal conditions have tended to focus on impairments (e.g., pain) and physical limitations (e.g., walking limitation). However, in recognition of the need to measure the full impact and wider influence of these conditions, there have been moves towards measuring the personal and social impact, captured by participation and social functioning (5–8). There is a growing interest in participation because the social consequences of musculoskeletal conditions (such as difficulty going shopping or visiting relatives) may be of more concern to patients than impairments (such as pain) or specific activity limitations (such as walking more than half a mile) (9). Participation is an important outcome measure for intervention studies and as a measure of success of health care and prevention programs, even if it is not the main target. For example, joint replacement or interventions to reduce pain may also enhance abilities to work or socialize with friends. Importantly, as many musculoskeletal conditions are long lasting, it is possible that even in the presence of ongoing signs (radiographic change), symptoms (pain), and activity limitation (walking limitation), participation can be maintained (5).

The purpose of this review was to assist the selection of an instrument to measure participation and social function in clinical practice or research studies in adult populations with musculoskeletal conditions. Following a comprehensive search of the published literature (the method is available from the authors), 6 instruments were identified that 1) had been developed to exclusively measure participation or social function in clinical practice or research, 2) were freely accessible and did not require purchase, and 3) have published evidence of sufficient psychometric testing to assess their applicability in (musculoskeletal) clinical practice or research. These instruments were the Impact on Participation and Autonomy (IPA) (10), Keele Assessment of Participation (KAP) (11), Participation Measure for Post-Acute Care (PM-PAC) (12), Participation Objective, Participation Subjective (POPS) (13), Rating of Perceived Participation (ROPP) (14), and The Participation Scale (15). All 6 have been developed as

Ross Wilkie, BSc, PhD, Joanne L. Jordan, BSc, MSc, MA, Sara Muller, BSc, MSc, PhD, Elaine Nicholls, BSc, MSc, Emma L. Healey, BSc, PhD, Danielle A. van der Windt, MSc, PhD: Keele University, Keele, Staffordshire, UK.

Address correspondence to Ross Wilkie, BSc, PhD, Arthritis Research UK Primary Care Centre, Primary Care Sciences, Keele University, Keele, Staffordshire, UK, ST5 5BG. E-mail: r.wilkie@cphc.keele.ac.uk.

Submitted for publication February 9, 2011; accepted in revised form May 9, 2011.

generic measures of participation. Four of the instruments (IPA, KAP, POPS, and ROPP) were developed to be administered as self-complete questionnaires. The Participation Scale and PM-PAC were designed to be administered as an interview and POPS can also be administered as an interview.

Each instrument is designed to measure participation in a different way; IPA measures choice and control (i.e., the possibility to do the things the way you want), KAP measures performance "as and when you want," PM-PAC measures limitation, POPS measures objective (i.e., frequency) and subjective (i.e., satisfaction) participation, ROPP measures the individual's perceived and desire to change participation, and The Participation Scale measures participation compared to a "peer norm." All of the instruments measure participation in mobility, self-care, domestic life, interpersonal interaction, and relationships, major life (e.g., work, education), and community and social life, except POPS, which does not measure aspects of self-care. The instruments contain a varying number of items (range 11–78 items); this is linked to the detail of participation measured (e.g., KAP contains the fewest items and measures participation broadly at domain level, POPS and ROPP contain the greatest number of items, and provide greater detail by measuring participation in specific life situations).

The evidence of psychometric properties is outlined in the Summary Table. The quality of psychometric testing varied across instruments. Tests have followed criteria and were linked to the intended measurement constructs. However, for 4 of the 6 instruments (KAP, POPPS, The Participation Scale, and ROPP) there is only 1 published paper that explored psychometric properties and most psychometric studies have been undertaken with small numbers. Tests for most instruments have focused on face and construct validity and reliability. Only IPA and PM-PAC have been tested specifically in musculoskeletal populations while the KAP has only been tested in the general population.

IPA, PM-PAC, POPS, and ROPP have been designed for use in clinical practice. All 4 instruments can indicate areas where restriction occurs and in addition the ROPP and POPS allow recipients to indicate areas of participation that they would like clinicians to focus on and aim to improve. KAP and The Participation Scale have not been tested for use in clinical practice as yet. IPA, KAP, and The Participation scale have been tested and applied in observational research studies (e.g., 5,15,16) but POPS, PM-PAC, and ROPP require further testing for such use. There is potential for IPA and The Participation Scale to be used in trials although further testing is required to assess their suitability to detect change (responsiveness).

IMPACT ON PARTICIPATION AND AUTONOMY (IPA)

Description

Purpose. The IPA was developed as a generic scale to measure person-perceived participation and autonomy, and for use in a wide range of populations (17). The original instrument was organized into 4 dimensions (social relationships, autonomy in self-care, mobility and leisure, and family role) and consists of 23 items. Updated versions have 5 domains (autonomy indoors, family role, autonomy outdoors, social relations, and work and educational opportunities), and 31 items plus 8 additional items to address problem experiences (10,18). English versions have 8 domains (31 items plus 8 items to address problem experiences) (19), and 5 domains (31 items plus 8 items to address problem experiences, plus 1 extra item [helping others]) (20).

Content. The items capture autonomy and participation. In the original instrument, following principal component analysis (PCA) and item reduction, 4 domains (23 items) are measured: social relations, autonomy in self-care, mobility and leisure, family role. Further development, again using PCA led to the 5 domains: autonomy indoors, family role, autonomy outdoors, social relations, work and education (10).

Number of items. 5 subscales; 31 items plus 8 problem experience items (10).

Response options/scale. 5-point Likert scale (excellent, very good, moderate, poor, very poor) (17). Later revised to very good, good, fair, poor, very poor (10). A 3-point Likert scale is used for problem experience (no, minor, severe).

Recall period for items. Current.

Endorsements. None.

Examples of use. Lund ML, Lexell J. Associations between perceptions of environmental barriers and participation in persons with late effects of polio. Scand J Occup Ther 2009;16:194–204.

Slim FJ, van Schie CH, Keukenkamp R, Faber WR, Nollet F. Effects of impairments on activities and participation in people affected by leprosy in The Netherlands. J Rehabil Med 2010;42:536–43.

Lund ML, Lexell J. Relationship between participation in life situations and life satisfaction in persons with late effects of polio. Disabil Rehabil 2009;31:1592–7.

Videler AJ, Beelen A, van Schaik IN, de Visser M, Nollet F. Limited upper limb functioning has impact on restrictions in participation and autonomy of patients with hereditary motor and sensory neuropathy. J Rehabil Med 2009;41:746–50.

Lund ML, Lexell J. Perceived participation in life situations in persons with late effects of polio. J Rehabil Med 2008;40:659–64.

Nieuwenhuijsen C, van der Laar Y, Donkervoort M, Nieuwstraten W, Roebroeck ME, Stam HJ. Unmet needs and health care utilization in young adults with cerebral palsy. Disabil Rehabil 2008;30:1254–62.

Bastiaenen CH, de Bie RA, Vlaeyen JW, Goossens ME, Leffers P, Wolters PM, et al. Long-term effectiveness and costs of a brief self-management intervention in women with pregnancy-related low back pain after delivery. BMC Pregnancy Childbirth 2008;8:19.

Verbunt JA, Seelen HA, Ramos FP, Michielsen BH, Wetzelaer WL, Moennekens M. Mental practice-based rehabilitation training to improve arm function and daily activity performance in stroke patients: a randomized clinical trial. BMC Neurology 2008;8:7.

Van de Port IG, van den Bos GA, Voorendt M, Kwakkel G, Lindeman E. Identification of risk factors related to perceived unmet demands in patients with chronic stroke. Disabil Rehabil 2007;29:1841–6.

Kos D, Duportail M, D'hooghe M, Nagels G, Kerckhofs E. Multidisciplinary fatigue management programme in multiple sclerosis: a randomized clinical trial. Mult Scler 2007;13:996–1003.

Lemmens J, I S M van Engelen E, Post MW, Beurskens AJ, Wolters PM, de Witte LP. Reproducibility and validity of the Dutch Life Habits Questionnaire (LIFE-H 3.0) in older adults. Clin Rehabil 2007;21:853–62.

Lund M, Nordlund A, Bernsping B, Lexell J. Perceived participation and problems in participation are determinants of life satisfaction in people with spinal cord injury. Disabil Rehabil 2007;29:1417–22.

Middelkamp W, Moulaert VR, Verbunt JA, van Heugten CM, Bakx WG, Wade DT. Life after survival: long-term daily life functioning and quality of life of patients with hypoxic brain injury as a result of a cardiac arrest. Clin Rehabil 2007;21:425–31.

Practical Application

How to obtain. Questionnaire is included in the appendix of original articles (10,17).

Method of administration. Self-administered questionnaire.

Scoring. Each item is scored from 1–5; item scores are summated within domains. It is unclear if these subscales include the problem experience items or if these are analyzed separately.

Score interpretation. Increasing scores indicate greater perceived participation. Score range varies depending on the number of items. No normative values available and no interpretation or cut points are given.

Respondent burden. Mean ± SD time to complete (for original 41 items) is 30 ± 15 minutes (17). Mean time to complete (for 39 items) is 19.3 minutes (19).

Administrative burden. No information on administrative burden. Training is not needed.

Translations/adaptations. Original is in Dutch (10,18), and there are translations into English (19) and (20). Rasch analysis shows that the IPA subscales and a 30-item IPA (minus 1 item) were invariant across Dutch and English cultures (20). Adaptations: generic questionnaire, no changes for musculoskeletal populations.

Psychometric Information

Method of development. Items were generated by experts (multidisciplinary group) based on International Classification of Impairment, Disability, and Handicap, and discussed with patients (small qualitative pilot study). Items were deleted if not considered relevant for at least 75% of patients, or if ambiguous (procedures only briefly described).

Acceptability. Low response rates, especially in patients with rheumatoid arthritis and fibromyalgia (42% and 37%, respectively) (10). Missing values ranged from 0–3% (17). The results of an interview study to assess acceptability were as follows: easy to complete (83%), no items embarrassing, a few items could be removed or added according to participants (19). No information is available on floor or ceiling effects.

Reliability. Internal consistency for the original version (17) by Cronbach's $\alpha = 0.86$ (social relations), 0.87 (self-care); 0.84 (family role), 0.85 (mobility and leisure); for the revised version (10) Cronbach's $\alpha = 0.86$ (social relations), 0.91 (autonomy indoors), 0.90 (family role), 0.81 (autonomy outdoors), 0.91 (work and education) (10). Test–retest reliability: weighted kappas for individual items: 0.56–0.90. Intraclass correlation coefficient range for domains is 0.83–0.91.

Validity. *Face and content validity.* IPA provides comprehensive coverage of participation domains, and nearly all subdomains (except some aspects of communication and religion). Results of the interview study (19) showed IPA was accurate for patient's situation (63%), design was good (74.3%), and it was relevant (74–97.1%) (19).

Structural validity. Factor structure determined by PCA (10,17).

Construct validity. Hypotheses tested to assess convergent (correlations 0.29—0.59) and discriminant validity (correlations 0.01–0.50) using London Handicap Scale, Short Form 36, and Sickness Impact Profile. Associations have not been presented for social functioning subscale of the Short Form 36 (10).

Ability to detect change. Anchor-based method. Before-after treatment and several transition indices (18). Standardized response means (SRMs) and area under the curve (AUC) indicate responsiveness to change of 3 dimensions; family role (SRM 0.8, AUC 0.8), autonomy outdoors (SRM 1.2, AUC 0.89), work and education (SRM 1.3, AUC 0.93). Results acceptable for autonomy indoors (SRM 0.4, AUC 0.62) and for social relations (SRM 0.1, AUC 0.5).

Quality of psychometric testing. The design of the questionnaire is acceptable, although there was limited input from patients in design. It is unclear how and by whom items were generated. Procedures for reducing items are briefly described, and there is no justification for choice of response options. Internal consistency is good (2 studies). Reliability is good (1 study). There is no information on measurement error. In terms of face and content validity, comprehensive measurement of participation and interview study shows IPA is relevant and acceptable (1 study). In terms of construct validity, structural validity is demonstrated by PCA (2 studies), but no confirmatory factor analysis has been conducted.

For the hypothesis testing, the correlations (convergent/discriminant) not all convincing (1 study). In cross-cultural validity, 1 study confirms the relevance and acceptability in English; 1 study using item-response theory shows that 30 of the items were invariant across English and The Netherlands populations. In terms of responsiveness, it demonstrated acceptable properties to allow measurement of responsiveness for 3 of 5 domains (1 study), unacceptable for 2 domains. It is unclear if this is caused by lack of change in population. There is no information on minimum important change (MIC) and on interpretability. Overall, the quality of psychometric testing has been good, but more evidence in musculoskeletal populations needed on construct validity, measurement error, responsiveness, MIC, and interpretability.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument has good face validity and provides comprehensive measurement of participation. It has been tested in patients with a wide range of conditions, in particular neuromuscular disease, spinal cord injuries, traumatic head injuries, multiple sclerosis, stroke, fibromyalgia, and rheumatoid arthritis. Relevant according to patients.

Caveats and cautions. Further psychometric testing is required particularly with respect to construct validity and responsiveness. Study populations have not been very large. Low response rate in 1 study. No evidence on MIC and interpretability.

Clinical usability. Administrative burden is small. Responder burden is small, although response rate was low in one study, and time to complete is quite long (20–30 minutes). Acceptable and relevant to patients (1 study). Unable to interpret scores at present.

Research usability. Interpretability as yet unclear. Cross-cultural validity established for The Netherlands and UK, not yet available or tested in other languages. Face validity, consistency, and reliability are good, supporting research use, but more evidence needed in other populations and settings to support construct validity, measurement error, responsiveness, MIC, and interpretability. Time to complete may limit usefulness in research projects measuring a wide range of concepts.

KEELE ASSESSMENT OF PARTICIPATION (KAP)

Description

Purpose. KAP was developed as a generic measure of person-perceived performance of participation "as and when you want." It is intended for use in adults in the general population. Published in 2005 by Wilkie et al (11), there are currently no updates or revisions.

Content. Items measure participation in the domains of mobility, self-care, domestic life, interpersonal interaction, major life, community, and social life.

Number of items. 11 items (15 including the screening questions).

Response options/scale. Each item has a 5-point adjective ordinal scale (all of the time, most of the time, some of the time, a little of the time, none of the time).

Recall period for items. 4 weeks.

Endorsements. None.

Examples of use. Wilkie R, Peat G, Thomas E, Croft PR. The prevalence of person-perceived participation restriction in community-dwelling older adults. Qual Life Res 2006;15:1471–9.

Wilkie R, Peat G, Thomas E, Croft PR. Factors associated with participation restriction in community-dwelling adults aged 50 years and over. Qual Life Res 2007;16: 1147–56.

Wilkie R, Peat G, Thomas E, Croft PR. Factors associated with restricted mobility outside the home in community-dwelling adults aged 50 years and over with knee pain: an example of use of the International Classification of Functioning to investigate participation restriction. Arthritis Rheum 2007;57:1381–9.

Wilkie R, Thomas E, Mottram S, Peat G, Croft P. Onset and persistence of person-perceived participation restriction in older adults: a 3-year follow-up study in the general population. Health Qual Life Outcomes 2008;6:92.

Practical Application

How to obtain. The instrument is part of the original article (11).

Method of administration. Self-administered questionnaire.

Scoring. Each item is dichotomized to define the presence (some, a little, none of the time) or absence (all or most of the time) of participation restriction. Total scores are calculated by summing the number of items where restriction occurs (0–11 items). A computer is unnecessary. There are no instructions for managing missing data.

Score interpretation. Scores range from 0–11 (where 0 = no restriction and 1 to 11 = any restriction). No restriction is reported by 53% of the general population.

Respondent burden. Completion takes 3 minutes. It is easy to complete, and there is a 98.2% completion rate.

Administrative burden. Not reported.

Translations/adaptations. At present, only in English. Data have been reported for knee pain populations.

Psychometric Information

Method of development. Items were generated by the authors for the International Classification of Functioning participation domains 4 to 9. No patients were involved in item selection, but assisted with formatting. There are no subscales of participation. Item-response theory has not been used.

Acceptability. It is easy to complete and understand the questions; there was a 98.2% completion rate. No information on missing data; 53% of responses had no restriction-ceiling effect.

Reliability. Internal consistency was not examined. For test–retest, the mean observed agreement over a 4-week period for dichotomized responses was 90%. Kappa values ranged from 0.20–0.71. Interrater reliability was not relevant. Minimal detectable change and SEM were not reported.

Validity. Cognitive and semistructured interviews found that the instrument comprehensively measured participation. KAP demonstrated high levels of agreement with the Reintegration to Normal Living index and Impact of Participation and Autonomy.

Ability to detect change. Responsiveness and minimum clinically important difference (MCID) have not been tested.

Quality of psychometric testing. The development of the questionnaire is acceptable. Patients were used to adapt items. Internal consistency was not tested. Tests of reliability lack detail. There is no information on measurement error. Face and content validity: interview studies with musculoskeletal patients demonstrated high levels of acceptance and that participation was comprehensively measured. For construct validity and hypotheses testing, hypotheses were prespecified but not overly specific. The levels of agreement assessed with relevant tools (Re-integration to Normal Living and the Impact on Participation and Autonomy). Responsiveness was not tested. There is no information on minimum important change, and more information is required on interpretability. Overall, there is a reasonable level of testing to allow measurement of participation at a single time point. More evidence is required on responsiveness.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument measures participation comprehensively and can be applied to the general population (generic measure; i.e., all musculoskeletal groups). It is brief and concise, with minimal responder burden. It has not been tested sufficiently to evaluate its appropriateness for evaluating interventions.

Caveats and cautions. Further psychometric testing is required for responsiveness, MCID and information required for how to deal with missing items.

Clinical usability. Although the instrument has been used in the general population, further testing in clinical populations is required. Responder burden is minimal. No assessment of administrative burden.

Research usability. Face and construct validity tests support use for cross-sectional research. There is no assessment of administrative burden, and the responder burden will not limit use.

PARTICIPATION MEASURE FOR POST-ACUTE CARE (PM-PAC)

Description

Purpose. PM-PAC was developed to measure participation outcomes of rehabilitation services provided in outpatient or home-care settings. It is a measure for community-based individuals, and was originally published by Gandek et al in 2007 (12).

Content. PM-PAC evaluates participation in 9 domains: mobility, role functioning, community, social, and civic life, domestic life/self-care, economic life, interpersonal relationships, communication, work, and education.

Number of items. 51 items.

Response options/scale. The PM-PAC contains 11 different response options: 1) Are you limited? (not at all/a little/some/quite a lot/completely); 2) How much are you limited? (not at all limited/a little/somewhat/very much/extremely limited/do not do this); 3) How much of the time? (all the time/most of the time/some of the time/a little of the time/none of the time); 4. How many days? (everyday/5–6 days/3–4 days/1–2 days/never); 5) Employment status? (working full-time/working part-time/unemployed but looking for work/unemployed and not looking for work/a homemaker/doing full- or part-time volunteer service/full-time student/employment trainee/vocational rehabilitation/retired/temporarily unable to work because of health or disability/completely unable to work due to health or disability); 6) Education? (yes/no but I would like to/no and I don't want to); 7) Describe social life (I do not have any difficulty doing things socially/I maintain my usual pattern of social activities/I am somewhat restricted in the amount and type of social activities I do/I am restricted in the amount and type of social activities I do/I do not see family and friends/I only see those who come to care for me); 8) How many times have you done things socially? (none/once/twice/3 times/more than 3 times); 9) Satisfaction (very satisfied/somewhat satisfied/neither satisfied nor dissatisfied/somewhat dissatisfied/very dissatisfied); 10) Number of close friends (0, 1, 2 to 4, 5 to 8, or 9+); 11) Drain on financial resources? (not at all/a little/somewhat/quite a lot/extremely).

Recall period for items. Past week or current status.

Endorsements. None.

Examples of use. Jette AM, Keysor J, Coster W, Ni P, Haley S. Beyond function: predicting participation in a rehabilitation cohort. Arch Physical Med Rehabil 2005;86:2087–94.

Keysor JJ, Jette AM, Coster W, Bettger JP, Haley SM. Association of environmental factors with levels of home and community participation in an adult rehabilitation cohort. Arch Physical Med Rehabil 2006;87:1566–75.

How to obtain. In appendix of reference 12.

Method of administration. Intended for self-report but was administered by interview during testing.

Scoring. Scoring instructions not given.

Score interpretation. Higher scores indicate greater participation and satisfaction. No normative values given.

Respondent burden. Assessed but results not reported.

Administrative burden. Not reported.

Translations/adaptations. Original in English, there is a French translation although not tested. PM-PAC has been further developed for administration using computer assisted testing (PM-PAC-CAT).

Psychometric Information

Method of development A literature search of existing tools was used to identify possible items, along with generation of items from a group of experts. The items in the PM-PAC were discussed with 4 focus groups of rehabilitation patients and were pilot tested in interviews with 8 individuals with disability. Feedback on items was obtained from 8 professionals in the rehabilitation field and items were modified accordingly. Subscales were generated to link with the International Classification of Functioning domains; however, initial psychometric analysis led to the allocation of some items to other domains. Potential items were selected by a group of experts. Item-response theory led to the allocation of some items to other domains.

Acceptability. Readability was not reported. Missing data do not appear to be common, although exact levels are not reported.

Reliability. For internal consistency, Cronbach's alpha for all scales ranged from 0.72–0.89. For test–retest reliability, the intraclass correlation coefficient range was 0.61–0.86. Mean difference scores ranged from −2.09 to 2.11 across the 7 scales (test–retest scores were not significantly different on average across the scales). Minimal detectable change and SEM were not tested.

Validity. Face validity was not tested. Confirmatory factor analysis indicated that there were 7 participation domains (work and education were not included). In terms of construct validity, scores differed significantly with diagnostic groups ($P < 0.001$) on all scales except domestic life. Mean scale scores generally reflected hypothesized patterns; those with more severe problems for mobility, role functioning, and community and social life. Number of days that respondents left their home was significantly related to mobility, role functioning, community, social and civic life, and domestic life but was not significantly related to the other scales.

Ability to detect change. Responsiveness and minimum clinically important difference were not tested.

Quality of psychometric testing. The development of PM-PAC is acceptable. There is a poor description of scoring, and no information on missing data or how to handle missing data. Internal consistency was adequately tested and is an acceptable level. Reliability was adequately tested and is an acceptable level. Face validity was not assessed.

Construct was examined with reference to impairment status. Further testing is required. Responsiveness was not assessed, and there is no information on minimum important change and interpretability. Overall, there is high quality of testing of internal consistency and reliability, but no other testing of other properties. Some of the remaining properties of the measure not tested here are dealt with in the development of PM-PAC-CAT, which is not freely obtainable.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. May be useful for clinical practice. Other adaptations, which are not freely available, show promise for computer-assisted testing, monitoring change, and testing interventions.

Caveats and cautions. Further testing is required to indicate levels of responder and administrative burden. Unsure how to score. Further psychometric testing is required particularly with relation to responsiveness and repeatability.

Clinical usability. Unable to interpret scores at present. Further assessment of administrative and responder burden would be useful.

Research usability. Unclear how to score the instrument. Could be used to measure participation in the 9 domains. More testing on responsiveness would be useful.

PARTICIPATION OBJECTIVE, PARTICIPATION SUBJECTIVE (POPS)

Description

Purpose. POPS was developed to measure participation (as defined by the International Classification of Functioning framework) and extend previous measures of community integration. It is a generic tool developed for any population. It was published in 2004 by Brown et al (13) and there are no known updates or revisions.

Content. The instrument measures participation in 5 domains: domestic life (8 activities), interpersonal interactions and relationships (8 activities), major life areas (3 activities), transportation (2 activities), and community/recreational and civic life (5 activities).

Number of items. For each of the 26 activities, there are 3 questions, giving a total of 78 items. The first measures frequency or duration of engagement (objective participation), the second measures how important engagement in the activity is, and the third refers to whether they would like to change their current level of engagement (subjective participation).

Response options/scale. For objective participation, response options are measured as amounts: percentage of the activity that an individual is responsible for (domestic life domain), number of hours per day, week, or month the activity is engaged in (major life areas domain), or the frequency of occurrence of the activity in a day, week, or month (all other domains). For subjective participation the importance of each of the 26 activities to well-being is coded using the response options most, very, moderate, little, or not important (scored 4–0), and any change to current level of engagement in that activity is rated as same, less, or more.

Recall period for items. Varies from current time up to 1 month.

Endorsements. None.

Examples of use. Mascialino G, Hirshson C, Egan M, Cantor J, Ashman T, Tsaousides T, et al. Objective and subjective assessment of long-term community integration in minority groups following traumatic brain injury. Neurorehabilitation 2009;24:29–36.

Cantor JB, Ashman T, Gordon W, Ginsberg A, Engmann C, Egan M, et al. Fatigue after traumatic brain injury and its impact on participation and quality of life. J Head Trauma Rehabil 2008;23:41–51.

Practical Application

How to obtain. The instrument is included in the appendix of the original article by Brown et al (13) and scoring instructions are available from the authors on request. No costs are reported.

Method of administration. Self-completion questionnaire or by interview/telephone.

Scoring. The scoring for the POPS is normalized using data from a sample of patients with traumatic brain injury and those with no disability.

Scoring objective participation. First, all hour and frequency items are converted to a single base, frequency, or duration per month. Standardized scores are then calculated by subtracting from each person's raw score for each item the mean score for the item for the combined population norm and dividing by the SD for the population. To control for outliers standardized scores are set to −3 and +3. Standardized scores are then weighted by a factor that was the average of the mean importance rating of the population so that all items do not have equal weighting (things that are done more have a greater weighting). The total score for objective participation is calculated as the average of the weighted standardized scores of the 26 items. Subscale scores can be calculated as the average standardized scores for the standardized samples.

Scoring subjective participation. For each of the 26 items, multiply the importance score by the satisfaction score, where a person who is wanting less or more was scored as −1, and his or her being satisfied with current level was scored as +1. Scores can range from +4, indicating a most important area of life that the person is engaging in at a satisfactory level, to −4, indicating an equally important area of life that the person wants to do either less of or more. The subjective participation total score is the mean of the 26 activities. A computer is necessary to score this tool. Further scoring instructions are available from the author and may give more detail on how missing data should be handled in the scoring algorithm.

Score interpretation. The score range for subjective participation is −4 to 4. The score range for objective participation is −3 to 3. Higher scores indicate greater participation and there are no recommended cut-offs. Normative values are not available for either scale, although data on the importance of the 26 items are given from a sample of patients with no disability. These data are included in the scoring algorithm for the tool.

Respondent burden. No information is give on item difficulty, although the instrument was successfully completed by 575 participants included in the psychometric evaluation.

Administrative burden. No information is given on time to administer the POPS. The POPS can be used as a self-report measure so difficulties for an administrator would be minimal and training would be minimal. The scoring of the tool would be done by computer and although time may be needed to set up the algorithms, once achieved, they should be able to be used with ease.

Translations/adaptations. Only available in English.

Psychometric Information

Method of development. The items originate from the Living After Traumatic Brain Injury instrument, which was drawn from a variety of existing instruments (mainly the Craig Handicap Assessment and Reporting Technique, Community Integration Questionnaire, Bigelow Quality of Life Questionnaire, and the Community Re-entry Questionnaire), although the process of selection is not reported. Additional items have been added, although there is no justification given for their inclusion. Patients were not used in development. Subscales were generated to fit with the International Classification of Functioning domains. Item-response theory was not used. Two separate subscales can be generated (PO and PS), using scoring algorithms available from the author. In addition, domain scores can be calculated for the 5 domains covered in the questionnaire.

Acceptability. The questionnaire is readable; however, no specific details are given on item-level missing data rates. In the scoring of the POPS, the authors offered an alternative scoring using deviation scores due to the skewed nature of some of the distributions. This suggests that floor and ceiling effects may be present in the data; however, floor and ceiling effects may only apply to the sample of brain injury patients included in this study.

Reliability. Internal consistency of scales ranged from 0.37–0.89, except for transportation, which was lower for both PO and PS. Test–retest reliability was completed on a subsample of patients with traumatic brain injury (n = 65). The time period for test–retest reliability was between 1 and 3 weeks. Both the PO and PS scores had good test–retest reliability (intraclass correlation coefficient 0.75 and 0.8, respectively); however, the test–retest reliability of the PO and PS scores for each domain varied between 0.28 and 0.89. No information is given on minimal detectable change or SEM.

Validity. *Face/content validity.* The items have been selected from other participation tools; however, little information is given on item face validity. It is not clear how the items were selected for the tool.

Construct validity. There is a weak correlation between PO and PS total scores (0.23 for mild traumatic brain injury, 0.21 for moderate-severe traumatic brain injury).

Four other measures of positive-negative affect or satisfaction judgments (Brain Injury Screening Questionnaire, Beck Depression Inventory II, Flanagan Quality of Life Scale, and Life 3) had stronger correlations with PS than PO total scores. Current age, injury severity, and years post–traumatic brain injury onset were not related to PO or PS total scores. Overall, there was moderate evidence to support the construct validity of the tool, although not all predefined hypotheses were supported in the data.

Ability to detect change. No evidence is given for the tool's responsiveness or minimum clinically important difference.

Quality of psychometric testing. The design of the questionnaire is adequate and is generated by "experts." The quality of assessment of the internal consistency is acceptable.

The study of reliability was of reasonable quality. There is no information on measurement error, and the face and content validity were not clearly reported. Construct validity was tested with regard to the hypothesis and relevant instruments. Responsiveness was not assessed, and there was no information on minimum important change. In terms of interpretability, objective and subjective participation are different constructs. Further work is required to interpret scores for each scale. Overall, further testing is required on missing data, face validity, and responsiveness and in musculoskeletal populations for use.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Measures both objective and subjective participation and allows patients to indicate the areas that they would like clinicians to target.

Caveats and cautions. The process used to select the 26 items for the POPS as a subset of those in the Living Life After Traumatic Brain Injury Study is unclear. The process for generating new items for POPS (e.g., in the transportation section) is also not clearly defined. No information is given regarding the interpretability of the scale units of the PO and PS scales. The scoring of the instrument is quite complex so it would be difficult to apply in clinical practice. Only pilot data are given in the assessment of the psychometric properties of the tool, so further work is needed in this area, especially around minimum clinically important difference and interpretation of scale scores.

Clinical usability. POPS scale has the potential to facilitate decisions on patient care, although these require further testing. Further testing of responder burden would be useful to guide clinical use. Clinical use may be limited by the scoring method, which requires a computer.

Research usability. Overall the psychometric data presented support use for research, although further psychometric testing is needed. The questionnaire can be self-administered; however, computer-generated scoring algorithms are needed to score the tool. Further testing is required on responsiveness.

RATING OF PERCEIVED PARTICIPATION (ROPP)

Description

Purpose. ROPP was developed to describe patient's perceived level of participation in the context of where they live and to direct rehabilitation interventions according to the patient's desire to change a particular domain. The ROPP was published in 2007 by Sandstrom and Lundin-Olsson (14). There are no revisions or updates.

Content. There are 22 statements about participation in different life situations, which are grouped into 9 different areas. Linked to each statement there are questions about participation level, satisfaction with participation, and if support is desired. Domains covered are: personal care, mobility, communication, social relationships, domestic life and caring for others, education, work and employment, economic life, and social and civic life.

Number of items. There are 22 statements with 3 questions per statement for a total of 66 items. A participation score (range from 0–88) can be derived using only the first question (participation level) for each of the 22 items.

Response options/scale. Perceived participation (not restricted, not applicable/mildly restricted/moderately restricted/very restricted/severely restricted); satisfaction (yes/no), need for support to change participation level (yes/no).

Recall period for items. Current situation.

Endorsements. None.

Examples of use. None, other than the original article (14).

Practical Application

How to obtain. Questionnaire is included in the original article (14).

Method of administration. Self-administered questionnaire.

Scoring. Each item is scored 0 (no restriction) to 4 (severe restriction). The scale score is calculated by adding the score for each item (0–88). Furthermore, for each item participants rated whether they were satisfied with their level of participation (yes/no) and whether they wanted support to change their level of participation (yes/no). At the end of the questionnaire, participants are asked to select which, out of the 9 domains given, are the 3 most important ones for changing the level of participation. No instructions for missing data.

Score interpretation. Increasing score means increasing participation restriction. There are no normative values.

Respondent burden. Takes 15–30 minutes to complete. When asked about the time to complete, 85% of responders were positive, 11% neutral, and 4% were negative.

Administrative burden. Not discussed although the instrument is self-report and expected to be minimal.

Translations/adaptations. English. Developed in neurologic population. A Swedish version may exist.

Psychometric Information

Method of development. Items were selected from the 9 domains of the International Classification of Functioning that were judged to be important for adults with mild to severe signs and symptoms of neurologic disease. Multidisciplinary staff then reviewed the proposed items for relevance, comprehensibility, and clarity. Items were adapted following pilot testing and cognitive interviews with patients. Subscales were generated to map to the International Classification of Functioning. Item-response theory was not used.

Acceptability. Questions were easy to understand (68% said the questions were understandable, 14% were neutral, and 18% were negative). Layout could be difficult to grasp, and changed for the final version of the questionnaire. 1.2% of data were missing. Floor effects occur and scores range from 0–60 with a mean ± SD 20.6 ± 1.8.

Reliability. *Internal consistency.* Cronbach's $\alpha = 0.90$ for the total score. Five of 6 domains (with 2 or more items $\alpha = 0.73$ to 0.89, and 0.50 for economic life).

Test–retest. Mean ± SD scores between test and retest was 1.13 ± 3.93. Difference between test and retest total scores calculated by intraclass correlation coefficient (1,1) and (3,1) were 0.97 (95% confidence interval [95% CI] 0.94–0.98) and 0.97 (95% CI 0.95–0.98), respectively. The within-subject SD was 2.9 points. The difference between 2 measurements for the same person was <7.9 points for 95% of pairs of observations.

Perceived participation, for 20 items weighted kappa was >0.70. Agreement on the selection of domains with a kappa above 0.80 in 7 domains and above 0.70 in the remaining 2 domains.

Validity. For face validity, 61% of patients were positive that the instrument measured participation, 23% were neutral, and 16% were negative; 85% of professionals were positive, 14% were neutral, and none were negative. Agreement between perceived participation and satisfaction, weighted kappa was at least 0.70 for all items. Agreement between perceived participation and desired support, weighted kappa for all reached at least 0.80.

Ability to detect change. Responsiveness and minimum clinically important difference were not tested.

Quality of psychometric testing. Internal consistency showed reasonable quality.

Reliability was of reasonable quality, measurement error was adequately tested, and there was reasonable quality study of face validity. Construct validity was only tested between perceived participation and satisfaction. Responsiveness was not tested; there is no information on minimum important change, and no information on interpretability. No item-response theory was used in development. Internal consistency and reliability were assessed, and both were of reasonable quality. Overall, there is a need for further testing of construct validity, interpretability, and responsiveness.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument measures all participation domains and allows people to indicate if they require interventions to improve their participation.

Caveats and cautions. Requires further psychometric testing of construct validity and responsiveness.

Clinical usability. This tool has the potential to aid clinical practice by allowing patients to indicate which participation areas they desire help with. Psychometric testing provides minimal support for use in clinical settings, and more is required. Administrative burden was not discussed. Responder burden may limit use, but appears minimal.

Research usability. Psychometric testing provides some support for use in observational studies, although more is required. Administrative burden not discussed. Responder burden may limit but appears minimal.

THE PARTICIPATION SCALE

Description

Purpose. The Participation Scale was developed to measure participation in comparison with a "peer norm." It was developed for people with leprosy or disability, spinal cord injury, and polio or other "stigmatized" conditions and originally published in 2006 by Brakel et al (15). There are no updates or revisions.

Content. The Participation Scale measures all participation International Classification of Functioning domains except for general tasks and demands.

Number of items. 18 items.

Response options/scale. There are 5 response options: 0 = no restriction; 1 = some restriction, but no problem; 2 = small problem; 3 = medium problem; 5 = large problem; 4 = no response.

Recall period for items. Not reported.

Endorsements. None.

Examples of use. Singh S, Sinha AK, Banerjee BG, Jaswal N. Participation level of the leprosy patients in society. Indian J Leprosy 2009;81:181–7.

Lesshafft H, Heukelbach J, Barbosa JC, Rieckmann N, Liesenfeld O, Feldmeier H. Perceived social restriction in leprosy-affected inhabitants of a former leprosy colony in northeast Brazil. Leprosy Rev 2010;81:69–78.

Wee J, Lysaght R. Factors affecting measures of activities and participation in persons with mobility impairment. Disability Rehabil 2009;31:1633–42.

Nicholls PG, Bakirtzief Z, Van Brakel WH, Das-Pattanaya RK, Raju MS, Norman G, et al. Risk factors for participation restriction in leprosy and development of a screening tool to identify individuals at risk. Leprosy Rev 2005;76:305–15.

Practical Application

How to obtain. The instrument is included in the original article (15). There is no cost to obtain it.

Method of administration. Interview-based instrument.

Scoring. Severity score <13 = 1, 13–22 = 2, 23–33 = 3, 34–53 = 4, >53 = 5, (where 1 = no significant restriction, 5 = extreme restriction). Not reported if a computer is necessary.

Score interpretation. Not clearly presented. Greater score equals greater restriction. Scored 1–5. Arbitrary severity categories provided. Normative values not reported.

Respondent burden. It takes 20 minutes to complete. Item difficulty not reported.

Administrative burden. 20 minutes to administer. Difficulty for administrator not reported. Training needed, but do not need to be a specialist.

Translations/adaptations. The Participation Scale is available in 6 other languages (for use in Brazil, Nepal, and India [Hindi, Bengali, Telugu and Tamil], although it is not clear how these have been tested).

Psychometric Information

Method of development. Items were generated from field work, observation, and focus groups. The role of patients is unclear. No subscales generated. Item-response theory was not used.

Acceptability. Positive feedback received, useful or very useful. No information on missing data. No specific information on floor or ceiling effects.

Reliability. Internal consistency by Cronbach's α = 0.92. Item to total correlation was 0.32–0.73. The intraclass correlation coefficients for interrater and intrarater reliability were 0.80 and 0.83, respectively. Minimal detectable change and SEM were not reported.

Validity. In terms of content validity, the instrument covers all International Classification of Functioning participation domains except for general tasks and demands.

Construct validity significantly correlated with an expert score, Eyes, Hands, Feet outcome measure and patient self-assessment.

Ability to detect change. Responsiveness was assessed using post–life-changing event. The measure was considered responsive if a statistically significant difference could be demonstrated between baseline and the post–life-changing event, if the minimum difference was at least 10 points. Minimum clinically important difference was not assessed.

Quality of psychometric testing. Internal consistency was calculated and showed an acceptable level, and an acceptable level of test–retest reliability was shown. There was no information on measurement error, and face and content validity were not evaluated. Construct validity was tested against expert opinion and the Eyes, Hands, Feet measurement tool (used in the populations with leprosy to measure limitation). Responsiveness was not tested, and there was no information on minimum important change and interpretability. No item-response theory was used in development. Overall, there is a need for further testing of construct validity, interpretability, and responsiveness.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument measures participation comprehensively.

Caveats and cautions. More detailed analysis needed to evaluate all psychometric properties. Has only been tested in people with stigmatized conditions/those with disability. No specific recall period. Not clear if evaluation has been done on English version.

Clinical usability. Takes 20 minutes to complete and needs a trained administrator. No assessment of administrative burden. Unclear if use is limited by responder burden. Requires more psychometric testing.

Research usability. Requires more psychometric testing.

DISCUSSION

This review has identified 6 instruments that have acceptable levels of evidence of face, content, and construct validity, and reliability to support their application to measure participation and social function at a single time point in clinical practice or research. Each instrument has individual strengths and weaknesses that may aid selection for use and additional reviews on measures of participation may provide further insight (21–23). At present, there is no need to develop more instruments; however, the current instruments should be tested more rigorously to support or refute their use. This suggestion is also relevant for other instruments not included in this review such as the Social Role Participation questionnaire (1) and the Valued Life Activities Questionnaire (24), which we are aware are freely available from their authors and are currently being tested further. None of the instruments highlighted in this review have been developed to measure participation or social function specifically in musculoskeletal populations, but are for use in other populations (e.g., general [11] or people with neurologic conditions [15]). This review has allowed an evaluation of the current status of participation/social function measurement and shows that there have been greater gains in developing methods in other fields from which the musculoskeletal community can benefit. The Participation Measure for Post-Acute Care (PM-PAC) has been soundly developed for administration using Computerized Adaptive Testing (i.e., PM-PAC-CAT) (25). Psychometric testing suggests that this tool has the ability to detect change and could be useful in clinical practice and intervention studies. However, the PM-PAC-CAT was not included in the review because it is not freely accessible.

We referred to the Consensus-Based Standards for the Selection of Health Measurement Instruments checklist to review the methodologic quality of psychometric studies, although the amount of testing that has been done limits judgment on the quality of, and support for using, these instruments (26,27). For all instruments, further testing is required in musculoskeletal populations to explore responsiveness and interpretation of scores. Longitudinal measurement of participation restriction needs to be explored further to ascertain the best method to capture change over time. The multidimensional and dynamic nature of participation presents a challenge that may be solved by focusing on individual aspects of participation. Issues of "response shift" may occur, with some dimensions becoming less important over time (e.g., restrictions in work), whereas others become increasingly relevant to people (e.g., self-care or mobility) (28). Testing of the unidimensionality of the Impact on Participation and Autonomy Scale and PM-PAC using factor analysis demonstrated that different areas of participation exist within these instruments. These should be assessed as different subscales, rather than be incorporated within a single scale, which may be assessing multiple concepts. The dimensionality of the other 4 instruments has not yet been investigated, and the identification of subscales and linked factors may help the development of methods to manage participation problems.

We took a stringent approach to selecting instruments for this review. Searching for instruments that capture participation/social function can unearth a large number of instruments because of similar and related concepts. It must be noted that there are a number of instruments that exist to provide detailed measurement of specific areas of participation (e.g., work), which are beyond the remit of this review. The field of participation and social function is evolving and our search identified a number of instruments that did not meet our criteria. For example, the social functioning scale of the Medical Outcomes Study Short Form 36 (29) has been used to measure participation (6). We excluded this instrument because of the costs associated with its use. It is also unclear whether the 2 items that make up this scale refer solely to participation, and they certainly do not cover the various dimensions covered by the instruments included in this review. Similarly the 2 items that constitute the Activities and Participation Questionnaire (30), which we also excluded from our review, capture "usual activities" and do not refer specifically to participation or social function. There are a number of other instruments which include some items that measure participation and are generic (e.g., Sickness Impact Profile [31]), have been designed to either measure a specific aspect of participation (e.g., Re-integration to Normal Living [32]), general functioning (e.g., WHODAS II [33]) or include other concepts, such as well-being (e.g., PIPP [34]). These instruments are applicable and may be very relevant in populations with musculoskeletal conditions, but do not exclusively assess participation or social function, and hence were not selected for our review.

Measuring participation offers the potential to capture the impact of musculoskeletal conditions in the context in which people live. However, its measurement is in its infancy. Our search has highlighted 6 instruments that could be used to measure participation/social function in adult musculoskeletal populations and have demonstrated minimally acceptable psychometric properties for use in clinical practice or research. None of these instruments as yet has been shown to have a clear advantage over the others. We stress that further psychometric testing is required to assess their measurement properties and applicability in a range of populations and settings, and therefore question the need to develop more instruments prior to such testing of existing instruments. This is important to support their use in practice and research for assessing the levels of restrictions and examining the success of interventions to maintain and improve participation and social function in those with musculoskeletal conditions.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Gignac MA, Backman CL, Davis AM, Lacaille D, Mattison CA, Montie P, et al. Understanding social role participation: what matters to people with arthritis? J Rheumatol 2008;35:1655–36.
2. Dijkers MP. Issues in the conceptualization and measurement of participation: an overview. Arch Phys Med Rehabil 2010;91 Suppl:S5–16.
3. World Health Organization. International Classification of Functioning, Disability and Health. Geneva: World Health Organization; 2001.
4. Jordan KP, Wilkie R, Muller S, Myers H, Nicholls E, Arthritis Research Campaign National Primary Care Centre. Measurement of change in function and disability in osteoarthritis: current approaches and future challenges. Curr Opin Rheumatol 2009;21:525–30.
5. Wilkie R, Peat G, Thomas E, Croft P. Factors associated with restricted mobility outside the home in community-dwelling adults aged 50 years and older with knee pain: an example of use of the International Classification of Functioning to investigate participation restriction. Arthritis Rheum 2007;57:1381–9.
6. Botha-Scheepers S, Watt I, Rosendaal FR, Breedveld FC, Hellio le Graverand MP, Kloppenburg M. Changes in outcome measures for impairment, activity limitation, and participation restriction over two years in osteoarthritis of the lower extremities. Arthritis Rheum 2008; 59:1750–5.
7. Machado GP, Gignac MA, Badley EM. Participation restrictions among older adults with osteoarthritis: a mediated model of physical symptoms, activity limitations, and depression. Arthritis Rheum 2008;59: 129–35.
8. Ayis S, Arden N, Doherty M, Pollard B, Johnston M, Dieppe P. Applying the impairment, activity limitation, and participation restriction constructs of the International Classification of Functioning model to osteoarthritis and low back pain trials: a reanalysis. J Rheumatol 2010; 37:1923–31.
9. Dunn JR, Cummins S. Placing health in context. Soc Sci Med 2007;65: 1821–4.
10. Cardol M, de Haan RJ, de Jong BA, van den Bos GA, de Groot IJ. Psychometric properties of the Impact on Participation and Autonomy questionnaire. Arch Phys Med Rehabil 2001;82:210–6.
11. Wilkie R, Peat G, Thomas E, Hooper H, Croft PR. The Keele Assessment of Participation: a new instrument to measure participation restriction in population studies: combined Qualitative and Quantitative Examination of its Psychometric Properties. Qual Life Res 2005;4:1889–99.
12. Gandek B, Sinclair SJ, Jette AM, Ware JE Jr. Development and initial psychometric evaluation of the participation measure for post-acute care (PM-PAC). Am J Phys Med Rehabil 2007;86:57–71.
13. Brown M, Dijkers MP, Gordon WA, Ashman T, Charatz H, Cheng Z. Participation objective, participation subjective: a measure of participation combining outsider and insider perspectives. J Head Trauma Rehabil 2004;19:459–81.
14. Sandstrom M, Lundin-Olsson L. Development and evaluation of a new questionnaire for rating perceived participation. Clin Rehabil 2007;21: 833–45.
15. Brakel WH, Anderson AM, Mutatkar RK, Bakirtzief Z, Nicholls PG, Raju MS, et al. The participation scale: measuring a key concept in public health. Disabil Rehabil 2006;28:193–203.
16. Lund ML, Lexell J. Associations between perceptions of environmental barriers and participation in persons with late effects of polio. Scand Journal Occup Ther 2009;16:194–204.
17. Cardol M, de Haan RJ, van den Bos GA, de Jong BA, de Groot IJ. The development of a handicap assessment questionnaire: the Impact on Participation and Autonomy (IPA). Clin Rehabil 1999;13:411–9.
18. Cardol M, Beelen A, van den Bos GA, de Jong BA, de Groot IJ, de Haan RJ. Responsiveness of the Impact on Participation and Autonomy questionnaire. Arch Phys Med Rehabil 2002;83:1524–9.
19. Vazirinejehad R, Lilley JM, Ward CD. The Impact on Participation and Autonomy: acceptability of the English version in a multiple sclerosis outpatient setting. Mult Scler 2003;9:612–5.
20. Kersten P, Cardol M, George S, Ward C, Sibley A, White B. Validity of the Impact on Participation and Autonomy Questionnaire: a comparison between two countries. Disabil Rehabil 2007;29:1502–9.
21. Magasi S, Post MW. A comparative review of contemporary participation measures' psychometric properties and content coverage. Arch Phys Med Rehabil 2010;91 Suppl:S17–28.
22. Noonan VK, Kopec JA, Noreau L, Singer J, Dvorak MF. A review of participation instruments based on the International Classification of Functioning, Disability and Health. Disabil Rehabil 2009;31:1883–901.
23. Resnik L, Plow MA. Measuring participation as defined by the International Classification of Functioning, Disability and Health: an evaluation of existing measures. Arch Phys Med Rehabil 2009;90:856–66.
24. Katz P, Morris A, Yelin EH. Prevalence and predictors of disability in valued life activities among individuals with RA. Ann Rheum Dis 2006;65:763–9.
25. Haley SM, Gandek B, Siebens H, Black-Schaffer RM, Sinclair SJ, Tao W, et al. Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation II: participation outcomes. Arch Phys Med Rehabil 2008;89:275–83.
26. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539–49.
27. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
28. Rapkin BD, Schwartz CE. Toward a theoretical model of quality-of-life appraisal: implications of findings from studies of response shift. Health Qual Life Outcomes 2004;2:14.
29. Ware JE, Snow KK, Kosinski M. SF-36 Health Survey: manual and interpretation guide. Lincoln (RI): QualityMetric Incorporated; 1993.
30. Li T, Wells G, Westhovens R, Tugwell P. Validation of a simple activity participation measure for rheumatoid arthritis clinical trials. Rheumatology (Oxford) 2009;48:170–5.
31. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787–805.
32. Wood-Dauphinee SL, Opzoomer MA, Williams JI, Marchand B, Spitzer WO. Assessment of global function: the Reintegration to Normal Living Index. Arch Phys Med Rehabil 1988;69:583–90.
33. Ustun TB, Chatterji S, Kostanjsek N, Rehm J, Kennedy C, Epping-Jordan J, et al. Developing the World Health Organization Disability Assessment Schedule 2.0. Bull World Health Organ 2010;88:815–23.
34. Pallant JF, Misajon R, Bennett E, Manderson L. Measuring the impact and distress of health problems from the individual's perspective: development of the Perceived Impact of Problem Profile (PIPP). Health Qual Life Outcomes 2006;4:36.

Summary Table for Measures of Social Function and Participation*

Scale	Purpose/ content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
IPA	23 items; measures choice and control of participation	Self-complete questionnaire	30 minutes	Minimal	Higher score = greater perceived participation restriction	Kappa for individual items 0.56–0.90; ICCs for domains 0.83–0.91	Comprehensive measure of participation; high levels of face validity; construct validity: associations with London Handicap Scale, and relevant components of the SF-36 and Sickness Impact Profile	Scales responsive to change: family role, (SRM 0.8, AUC 0.8), autonomy outdoors (SRM 1.2, AUC 0.9), work and education (SRM 1.3, AUC 0.9); scales less able to detect change: autonomy indoors (SRM 0.4, AUC 0.6), social relations (SRM 0.1, AUC 0.5)	Acceptable levels of validity and applicable in any population	Further testing required for construct validity and responsiveness
KAP	11 items; measures person-perceived performance in participation tasks	Self-complete questionnaire	3 minutes; 98.2% completion rate	Minimal	Range 0–11, higher score = more restrictions	Mean observed agreement over a 4-week period for dichotomized responses was 90%; kappa for individual items 0.20–0.71	Comprehensive measure of participation; high levels of face validity; construct validity: associations with IPA and Re-integration to Normal Living Index	Not tested	Brief, concise, and comprehensive; applicable in general adult populations	Further testing required for responsiveness; information required on managing missing data
PM-PAC	51 items; measures limitation in and satisfaction with participation	Interview	Not clearly reported	Not reported	Higher score = greater participation and satisfaction	ICC range 0.61–0.86; mean difference scores ranged from –2.09 to 2.11 across 7 scales	Construct validity: scores differed significantly with diagnostic groups ($P < 0.001$) on all scales except domestic life. (i.e., lower mean scores with greater severity of condition)	Not tested; other versions (PM-PAC-CAT) have demonstrated an ability to detect change	May be useful for clinical practice, but other adaptations not freely available (PM-PAC-CAT), show promise for CAT, monitoring change and testing interventions	Scoring is unclear; further testing is required for construct validity and responsiveness
POPS	78 items; measures objective and subjective participation in 26 activities	Self-complete or interview	No information	Minimal	Range for subjective participation −4 to 4; range for objective participation −3 to 3; higher scores = greater participation	Test–retest reliability $ICC_{PO} = 0.75$, $ICC_{PS} = 0.80$	Construct validity: association with BISQ, BDI II, Flanagan QOLS, Life 3; discordance between participation objective and participation subjective	Not tested	Captures objective and subjective participation; could be useful in clinical practice as allows patients to indicate areas of participation that they would like clinicians to target	Further testing required in musculoskeletal populations and for responsiveness; information required on managing missing data
ROPP	66 items; measures perceived level, satisfaction, and need for support to change the level of participation in 22 activities	Self-complete questionnaire	15–30 minutes	Minimal	Range 0–88; higher score = greater participation restriction	Internal consistency: Cronbach's α 0.50–0.89; mean ± SD difference between test–retest scores was 1.13 ± SD 3.93	Construct validity: associations with satisfaction and desired support	Not tested	Potentially useful in clinical practice as allows respondents to indicate the requirement of interventions to improve their participation	Further testing required for construct validity and responsiveness
The Participation Scale	18 items; compare an individual's participation to a peer norm	Interview	20 minutes to complete; item difficulty not reported	20 minutes to administer	Range 1–5; higher score = greater restriction; arbitrary severity categories provided	High level of internal consistency and reliability	Comprehensive measure of participation; reasonable level of construct validity	Found to be responsive	Measures participation comprehensively; more analysis needed	Further testing required; no specific recall period

* IPA = Impact on Participation and Autonomy; ICCs = intraclass correlation coefficients; SF-36 = Short Form 36; SRM = standardized response mean; AUC = area under the curve; KAP = Keele Assessment of Participation; PM-PAC = Participation Measure for Post-Acute Care; CAT = computer-assisted testing; POPS = Participation Objective, Participation Subjective; PO = objective participation; PS = subjective participation; BISQ = Brain Injury Screening Questionnaire; BDI II = Beck Depression Inventory II; Flanagan QOLS = Flanagan Quality of Life Scale; ROPP = Rating of Perceived Participation.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S86–S97
DOI 10.1002/acr.20531

Measures of Fibromyalgia

Fibromyalgia Impact Questionnaire (FIQ), Brief Pain Inventory (BPI), Multidimensional Fatigue Inventory (MFI-20), Medical Outcomes Study (MOS) Sleep Scale, and Multiple Ability Self-Report Questionnaire (MASQ)

DAVID A. WILLIAMS[1] AND LESLEY M. ARNOLD[2]

INTRODUCTION

The assessment of fibromyalgia (FM) is challenging because there are no biomarkers for this condition. Clinicians must rely upon patient-reported symptoms in order to understand the complexities of this condition. While in 1990, the American College of Rheumatology (ACR) developed research classification criteria involving tender point counts, it has only been within the past year that the ACR proposed clinical diagnostic criteria (1). Historically, many symptoms have been thought to be associated with FM. In order to narrow the field to those symptoms with the greatest clinical relevance, a working group within Outcome Measures in Rheumatology (OMERACT) conducted several Delphi exercises within both patients and clinicians to obtain consensus regarding which domains should be assessed in clinical trials for FM (2,3). The instruments to be reviewed herein reflect the clinically relevant domains defined by this OMERACT working group.

A wide variety of instruments have been used to index the OMERACT domains for FM. Many of the instruments were developed for use generically or have been borrowed from other clinical populations. In recent phase II and III clinical trials of medications for FM, wide variation was observed in the selection of domain indices (Table 1). While many of these measures are reviewed elsewhere in this special issue, we have selected a representative measure from each of the following domains of relevance: pain (Brief Pain Inventory), fatigue (Multidimensional Fatigue Inventory), sleep disturbance (Medical Outcomes Study Sleep Scale), and cognitive dysfunction (Multiple Ability Self-Report Questionnaire). Mood and functional status are also important domains for FM; however, the instruments most commonly used to assess these domains are reviewed elsewhere in this special issue and will not be repeated here (e.g., mood [Hospital Anxiety and Depression Scale] and functional status [Short Form 36]). Recent work in the development of responder indices suggests that either these specific instruments or other measurement tools from within the same domain can be used to differentiate responders from nonresponders in clinical treatment trials for FM (4). The precision by which these domains will be able to be assessed in the future is likely to be enhanced as newer measurements are being developed using either classic test construction methods or methods such as item response theory and computer adaptive testing, as is being done in the National Institutes of Health Patient-Reported Outcomes Measurement System (5).

FIBROMYALGIA IMPACT QUESTIONNAIRE (FIQ)

Description

Purpose. The FIQ was developed in the late 1980s by clinicians at Oregon Health & Science University (OHSU) to assess the total spectrum of problems related to fibromyalgia (FM) and associated responses to therapy (6). The FIQ was first published in 1991 (7) and modified in both 1997 and 2002 to refine items and to clarify the scoring system (6). The FIQ was revised in 2009 (FIQR) to better reflect current understanding of FM and to address limitations of the original FIQ while retaining its essential properties (8).

Supported in part by the National Institute of Arthritis and Musculoskeletal and Skin Diseases/NIH (grants U01-AR55069 and R01-AR053207).

[1]David A. Williams, PhD: University of Michigan, Ann Arbor; [2]Lesley M. Arnold, MD: University of Cincinnati College of Medicine, Cincinnati, Ohio.

Dr. Williams has received consultancy fees, speaking fees, and/or honoraria (less than $10,000 each) from Bristol-Meyers Squibb, Forest Pharmaceuticals, Jazz Pharmaceuticals, and Lilly. Dr. Arnold has received consultancy fees (less than $10,000 each) from AstraZeneca, Cypress Bioscience, Eli Lilly, Forest Pharmaceuticals, Takeda, Sanofi-Aventis, Grünenthal Group, Wyeth, and Johnson & Johnson, and (more than $10,000) from Pfizer, and has received grants from Boehringer Ingelheim, Cypress Bioscience, Eli Lilly, Forest Pharmaceuticals, Novartis, and Pfizer.

Address correspondence to David A. Williams, PhD, Chronic Pain & Fatigue Research Center, University of Michigan, 24 Frank Lloyd Wright Drive, Lobby M, Ann Arbor, MI 48106. E-mail: daveawms@umich.edu.

Submitted for publication February 28, 2011; accepted in revised form May 10, 2011.

Table 1. Outcome measures in fibromyalgia trials of Food and Drug Administration–approved medications

Fibromyalgia domain	Outcome measure
Pain	Visual analog scale (daily diary)
	Numeric rating scale (0–10) (daily diary)
	Fibromyalgia Impact Questionnaire pain (0–10)
	Brief Pain Inventory pain severity scores (0–10)
	Short Form 36 bodily pain
Tenderness	Dolorimetry (tender point threshold)
Fatigue	Visual analog scale (0–100) (daily diary)
	Fibromyalgia Impact Questionnaire fatigue (0–10)
	Short Form 36 vitality
	Multidimensional Fatigue Inventory
	Multidimensional Assessment of Fatigue
Sleep	Numeric rating scale (0–10) daily diary of sleep quality
	Fibromyalgia Impact Questionnaire morning rested feelings (0–10)
	Medical Outcomes Study sleep scale
Depression	Beck Depression Inventory
	Hamilton Depression Rating Scale
	Fibromyalgia Impact Questionnaire depression (0–10)
	Hospital Anxiety and Depression Scale depression
Anxiety	Fibromyalgia Impact Questionnaire anxiety
	Hospital Anxiety and Depression Scale anxiety
Cognition	Multiple Abilities Self-Report Questionnaire
Stiffness	Fibromyalgia Impact Questionnaire stiffness (0–10)
Physical function	Short Form 36 physical function
	Fibromyalgia Impact Questionnaire physical function

Content. The original FIQ (1991) covered 3 domains: function, overall impact, and symptoms. The function domain contained 10 physical functioning items related to the ability to perform large muscle tasks, including the ability to do shopping, do laundry, prepare meals, wash dishes by hand, vacuum a rug, make beds, walk several blocks, visit friends or relatives, do yard work, and drive a car. The overall impact domain contained 2 items asking about the number of days individuals felt well and the number of days they were unable to work because of FM symptoms. The symptoms domain contained 7 items using 10-cm visual analog scales on which patients rate work difficulties, pain, fatigue, morning tiredness, stiffness, anxiety, and depression. The 1997 version modified items about "work" to include "housework," and a new item about "climbing stairs" was added to the functioning domain. Finally, the 1997 version added hash marks (i.e., vertical lines) every 1 cm to the formatting of all visual analog scales. The 2009 FIQR has the same 3 domains as the original FIQ (function, overall impact, and symptoms), but differs in several ways. First, the physical functioning domain was reduced to 9 items and modified to reflect a better balance between large-muscle activities in the upper and lower extremities, and that would have less sex and ethnicity bias. The physical functioning items include the ability to brush or comb hair; walk continuously for 20 minutes; prepare a homemade meal; vacuum, scrub, or sweep floors; lift and carry a bag full of groceries; climb 1 flight of stairs; sit in a chair for 45 minutes; and go shopping for groceries. The overall impact domain was completely revised to reflect the overall impact of FM on functional ability and the overall impact of FM on the perception of reduced function. The symptom domain retained items on pain, fatigue, morning tiredness, stiffness, anxiety, and depression and added 4 additional items on tenderness, memory, balance, and environmental sensitivity.

Number of items. The original FIQ (1991) had 19 items capturing 3 domains. The 1997 version of the FIQ retained the same domains but added an additional item for a total of 20 items. In the 2009 FIQR, the first domain (physical function) has 9 items, the second domain (overall impact) has 2 items, and the third domain (symptoms) has 10 items for a total of 21 items.

Response options/scale. The physical functioning items in the 1991 and 1997 versions of the FIQ are rated on a 0–3 scale that best reflects the patient's ability to do the activity (0 = always, 1 = most, 2 = occasionally, 3 = never). The overall impact items are rated on a 0–7 scale for the number of days the patient felt well and the number of days the patient missed work, respectively. The symptom items are visual analog scales (0–10 cm), with higher numbers indicating greater symptomatology. All of the items in the 2009 FIQR are 0–10 numeric rating scales using 11 boxes, with higher numbers reflecting greater severity.

Recall period for items. The recall period is over the past week.

Endorsements/examples of use. Since 1991, the FIQ has been one of the most frequently used assessment tools in the evaluation of FM, and has been particularly useful as an outcome measure in FM clinical trials. The FIQ has been cited in over 300 articles between 1991 and 2010 (see URL: www.myalgia.com/FIQ/FIQ_REFS_2010.htm for a complete listing of article abstracts). The use of the FIQR in clinical studies has not yet been published.

Practical Application

How to obtain. The FIQ and the FIQR are free for academic and clinical use. An online license to use the FIQ is available by registering at URL: www.myalgia.com/FIQ/FIQ_academic_agreement.htm. The original FIQ is published in reference (7). The 1997 version with the 2002 scoring revision was published in 2005 (6) and is also available at URL: www.myalgia.com/FIQ/FIQ_B.htm. The FIQR is available at this same web site and was published in 2009 (8).

Method of administration. The FIQ and FIQR are administered as self-report questionnaires.

Scoring. The 1991 and 1997 FIQ versions have similar scoring. The final scores for each item of the FIQ should range from 0 (no impairment) to 10 (maximum impairment). The physical functioning items are rated on a 4-point Likert-type scale. Raw scores on each question can range from 0 (always) to 3 (never). Because some patients may not do some of the tasks listed, they are given the option of deleting questions from scoring. The scores for the items that the patient has rated are summed and divided by the number of questions answered. An average raw score between 0 and 3 is obtained. This value is then multiplied by 3.33. The first impact item that asks the number of days in the past week the patient felt well is reverse scored so that a higher number indicates impairment. Raw scores range from 0–7 and are then multiplied by 1.43. The second impact item is scored as the number of days the patient was unable to do regular work activities. Raw scores range from 0–7 and are then multiplied by 1.43. Symptom items are visual analog scales. In the 1991 version, the items are scored in number of cm from 0–10. Because the 1997 version added hash marks to all of the visual analog scales, these items are scored in numerical increments from 0–10, allowing scores to include 0.5 if the patient marks the space between 2 vertical lines. In the 1991 version, patients were instructed to cross out items 3 and 4 if they did not work. Therefore, the total maximum FIQ score was reduced from 100 to 80. With the 1997 revision in which questions 3 and 4 were modified to include housework, the total FIQ scores should always range from 0–100. In 2002, a modification of the scoring was recommended to address incomplete data. In order to maintain homogeneity on a 0–100 continuum, the final score is to be adjusted to reflect a final maximum score of 100. For example, if a patient missed 2 questions, the total recorded score should be adjusted by a factor of 10/8. The FIQR has 21 individual items and all items are based on an 11-point numeric rating scale of 0–10, with 10 being the "worst." The summed score for the function domain, which contains 9 items (range 0–90) is divided by 3; the summed score for overall impact, which contains 2 items (range 0–20) is not changed; and the summed score for symptoms, which contains 10 items (range 0–100) is divided by 2. As in the FIQ, the total maximum score for the FIQR is 100. The weighting of the 3 domains is different from the FIQ in that function accounts for 30% of the total score as opposed to 10% in the FIQ, the symptom domain makes up 50% of the score instead of 70% in the FIQ, and the overall impact domain remains the same as the FIQ at 20% (8).

Score interpretation. The final scores for each of the FIQ and FIQR items range from 0 (no impairment) to 10 (maximum impairment). The total maximum score for both the FIQ and the FIQR is 100, which represents the maximum impact of FM on the patient.

Respondent burden. It takes approximately 3–5 minutes to complete the FIQ. The FIQR is estimated to take just over 1 minute to complete.

Administrative burden. The FIQ and FIQR are easily administered by handing the questionnaires to the participant. The scales include simple instructions for the respondents. No formal training is required for the FIQ or FIQR. Scoring is relatively simple for both the FIQ and the FIQR but the use of numeric rating scoring for all of the FIQR items further simplifies the scoring and allows for use of electronic versions of the FIQR that can be administered online as was done in the validation study (8).

Translations/adaptations. The FIQ has been translated from English into 12 languages: Czech (Czech Republic), Dutch (The Netherlands), French (France and Canada), German (Germany), Hebrew (Israel), Italian (Italy), Korean (Korea), Polish (Poland), Romanian (Romania), Spanish (Argentina and Spain), Swedish (Sweden), Turkish (Turkey; see URL: www.myalgia.com/FIQ/FIQ_B.htm for more information on translations).

Psychometric Information

Method of development. The initial version of the FIQ was based on an intake questionnaire used by the OHSU rheumatology clinic and informal discussions with patients with FM. This FIQ was mailed at weekly intervals for a total of 6 weeks to 64 women with FM, along with the Arthritis Impact Measurement Scale (AIMS). A second group of 25 women with FM attending the OHSU Fibromyalgia Treatment Clinic completed the FIQ as part of their routine clinical evaluation. The construct validity, test–retest reliability, and content relevance of the FIQ were assessed in these 2 groups of patients (6,7). The FIQR was based on previous experience with the FIQ and patients' evaluation of important symptoms (8). The new questionnaire was tested in a focus group of 10 female patients with FM. Following discussions among the patients and investigators, agreement was reached on the final version of the FIQR. The FIQR was then tested in an online survey that was completed by patients with FM, rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), or major depressive disorder (MDD), and healthy controls. The participants also completed the original FIQ and the 36-item Short Form Health Survey (SF-36).

Acceptability. The FIQ was originally developed to assess the current health status of women with FM, and may therefore have a sex bias, particularly in the functional items in which several of these questions relate to activities that are more likely to be performed by women. The functional questions were intended for a relatively affluent patient who was assumed to have possession of a car, a vacuum cleaner, and a washing machine and may therefore not generalize to all patients with FM. The FIQ also has problems related to the deletion of physical function items deemed "not applicable" by the respondent, which may result in an underestimation of functional severity. Some patients report difficulty understanding the scoring of the physical function questions and note that the questions do not allow them to rate the degree of difficulty in performing the activity. For example, a patient may report that they were "always" able to do shopping even though it took a great deal of time and effort to complete the task. The FIQ functional items are oriented toward high levels of disability, resulting in a potential floor effect. For example, in one study, 12% of patients scored a 0 on the FIQ physical function score (i.e., no dysfunction) (9). The FIQR was developed to correct some of the problems with the FIQ. In particular, the physical functioning items were revised to have less sex and ethnicity bias than the FIQ and to improve the ease of scoring the functional activities on a 0–10 scale ranging from “no difficulty” to “very difficult” (8).

Reliability. In the original 1991 study to evaluate the FIQ, the test–retest reliability (Pearson’s r) was assessed by the weekly recording of data over 6 weeks. The reliability ranged from 0.56 on the pain score to 0.95 for physical function (7). The internal consistency (Cronbach’s alpha) was not reported in the original analysis. The Cronbach’s alpha for the FIQR was 0.95, with item-total correlations ranging from 0.56–0.93. Test–retest reliability was not determined for the FIQR (8).

Validity. The content validity of the original FIQ was assessed from an analysis of missing data for each item. Missing data from the physical functioning items were limited to 11% of patients who did not do dishes by hand and 20% who did no yard work. Because many patients were not working outside the home, the 2 work items of the original FIQ were not relevant for 38% of the patients (6,7). In the validation study of the FIQR, patient suggestions about content and wording of the instrument during the focus group meeting contributed to the face validity of the final version of the FIQR. Content validity of the FIQR was suggested by strong correlation between the FIQR and the SF-36. For example, the FIQR function domain was most highly correlated with the SF-36 physical functioning subscale (8). The construct validity of the 1991 FIQ was determined by measuring the correlation of the FIQ individual items with the AIMS. The FIQ physical functioning items had a significant correlation ($r = 0.67$) with the AIMS lower-extremity physical function component score. The pain, depression, and anxiety items of the FIQ showed significant correlations with the corresponding AIMS scales (0.69, 0.73, and 0.76, respectively). The AIMS visual analog of syndrome impact correlated least robustly with the FIQ items, the highest correlation being with pain ($r = 0.48$). Item correlations with the AIMS syndrome activity question tended to be higher, ranging from 0.28–0.83. A principal components analysis yielded 5 factors. The 10 physical functioning questions loaded on the first factor with component loading ranging from 0.50 to 0.95. Factor 2 consisted of work difficulty, feeling good, pain, fatigue, rest, and stiffness. Anxiety, depression, and days of work missed all loaded on separate factors (6,7). Convergent validity was assessed by comparing the FIQR to both the SF-36 and the FIQ. The 3 domains of the FIQR and the associated individual items correlated closely with the corresponding subscales on the SF-36. Each of the 3 FIQR domains was also highly correlated with the total FIQR score. There was a strong correlation (0.88) between the FIQR and the FIQ, suggesting that the questionnaires are capturing similar information about the impact of FM. The mean total score of the FIQR was ~4 points lower than the mean FIQ total score, which was attributed to the change of the weighting in the FIQR scoring (8). Each of the 3 FIQR domains predicted unique variance in SF-36 domains, providing evidence for discriminant validity. Discriminant validity was also evaluated by comparing the FIQR total scores in patients with FM with the scores in healthy controls, patients with RA or SLE, and patients with MDD. The FM FIQR scores were significantly higher than in the other 3 groups (8).

Ability to detect change. The FIQ has been most commonly used as an outcome measure in treatment trials and, in general, has demonstrated an ability to detect clinical change (6). The FIQ total score was also included as an outcome measure in trials of the 3 US Food and Drug Administration–approved medications for FM, pregabalin, duloxetine, and milnacipran (10–12). For example, in a pooled analysis of 4 placebo-controlled, double-blind studies of duloxetine in FM, the total FIQ scores improved significantly in the duloxetine groups compared with placebo, with a mean (SE) reduction of 12.62 (0.61) in the duloxetine patients compared with a mean (SE) reduction of 8.2 (0.69) in the placebo group ($P < 0.001$) (13). A recent study suggested that a 14% change or an absolute change of 8.1 (95% confidence interval 7.6–8.5) in the FIQ total score represented a clinically meaningful change in FM status (i.e., minimum clinically important difference). The minimum clinically important difference was determined by calculating the percentage change in the FIQ total score from baseline and linking this to each patient’s global assessment of change score (14).

References. The validation of the original FIQ is published in an article by Burckhardt et al (7). A review of the development, operating characteristics, and uses of the FIQ was done by Bennett (6) and the validation study of the FIQR is found in the Bennett et al publication in 2009 (8).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. FM is associated with multiple symptoms and functional impairment. The FIQ and FIQR are useful assessment tools in FM because they evaluate the total spectrum of problems related to FM, including functional impairment, overall impact, and FM-related symptoms. The FIQ total score has proved to be a useful outcome measure in key clinical trials of FM.

Caveats and cautions. The FIQ functional items are oriented toward high levels of disability, resulting in a possible floor effect. Because the FIQ was originally developed in a patient population of relatively affluent women, there is a potential problem with sex and ethnicity bias. Although the individual domains and/or items on the FIQ were not originally intended to be used in isolation, some recent studies have reported single-item or domain scores from this instrument. The internal consistency (Cronbach’s alpha) was not reported in the original analysis of the FIQ. The FIQR was designed to correct some of the problems with the FIQ, but has not yet been tested in the context of clinical trials. Test–retest reliability was not determined for the FIQR.

Clinical usability. The FIQ and FIQR are brief, self-report questionnaires that assess the impact of FM on patients. The FIQ has most commonly been used in clinical studies, but has the potential for use in the clinical setting to monitor patients' response to treatment over time.

Research usability. The FIQ has been used in large-scale clinical trials of therapeutics for FM, supporting its ability to assess and detect change in FM.

BRIEF PAIN INVENTORY (BPI)

Description

Purpose. The BPI was designed to measure multiple clinically relevant aspects of pain such as pain intensity and interference from pain in cancer populations (15). The BPI was originally called the Wisconsin Brief Pain Questionnaire (16). Subsequently, support for its valid use in noncancer populations such as musculoskeletal, neuropathic, and other central pain conditions has been established (17,18). There are 2 versions; the short version is the most commonly used and is often included in the context of clinical trials. This is the version that possesses most foreign language translations. A longer, less frequently used version is available that includes more pain descriptors and may have clinical utility; however, the developers recommend the short form for most applications. Only the shorter form will be considered here.

Content. The BPI assesses for the presence of pain, pain intensity (i.e., worse, least, average, current), and functional interference from pain (i.e., activity, mood, walking ability, normal work, relations with others, sleep, life enjoyment). It also catalogs the types of pain medications being used, the percentage of pain relief obtained from medications, and assesses the distribution of pain via a body map.

Number of items. The BPI contains a total of 15 items.

Response options/scale. The BPI uses a mixture of item types. Item 1 querying about the presence of pain is a dichotomous "yes," "no." Item 2, the body map, asks that areas of pain be shaded and an "X" placed on the body region that hurts the most. Items 3–6 (intensity items) utilize a 0 (no pain) to 10 (pain as bad as you can imagine) 11-point rating scale. Item 7 is an open-ended response to list pain medications. Item 8 (percentage of pain relief) ranges from 0% (no relief) to 100% (complete relief). Item 9 (a–g) inquires about interference using an 11-point numeric rating scale. Each item ranges between 0 (does not interfere) and 10 (completely interferes).

Recall period for items. The time frame for the BPI is typically based upon "the past week" but some versions allow for the past 24 hours.

Endorsements/examples of use. The BPI is widely used in clinical trials for pain and in pain research generally. It is one of the instruments recommended by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials group (19) for inclusion in any clinical trial evaluating pain.

Practical Application

How to obtain. The BPI is available through the following address: The Department of Symptom Research, Attn: Assessment Tools, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Unit 1450, Houston, TX 77030. Phone: 713-745-3805. The BPI is available free of charge for nonfunded academic research. For funded academic research there is a charge per project (e.g., $300) and a charge for commercial research (e.g., $800 per project).

Method of administration. The BPI can be administered as a self-report questionnaire or as an interview.

Scoring. While some of the items represent single-item values, pain intensity, indexed by the Pain Severity Score, is calculated by obtaining the mean of the 4 pain intensity items. The Pain Interference Score is obtained by calculating the mean of the 7 interference items.

Score interpretation. The Pain Severity Score has a maximum value of 10 (i.e., "pain as bad as you can imagine") and a minimum value of 0 (i.e., "no pain"). The Pain Interference Scale similarly has a maximum value of 10 (i.e., "completely interferes") to 0 (i.e., "does not interfere"). The BPI is easily scored by hand.

Respondent burden. It takes approximately 5 minutes to complete the BPI.

Administrative burden. The BPI is easily administered by handing the questionnaire to the participant or by asking each question verbally. Scoring is accomplished by calculating 2 means, which can be done in <5 minutes.

Translations/adaptations. Validated translations are available for the following languages: English, Spanish, Italian, Russian, Norwegian, Greek, German, Japanese, Chinese, Arabic, Bulgarian, Cebuano, Croatian, Czech, Filipino, French, Hindi, Korean, Malay, Slovak, Slovenian, and Thai.

Psychometric Information

Method of development. Prior to the development of the BPI, there was no specific instrument designed to the intensity and impact of cancer pain that was brief and that could be administered repeatedly over time to monitor the effects of treatment. Existing measures at the time (e.g., the McGill Pain Questionnaire) were developed for noncancer pain. Based upon patient interviews, it was discovered that existing questionnaires were too ambiguous, irrelevant, or too lengthy for the assessment of cancer pain. The questionnaire was developed in accordance with the best guidelines for test construction available at the time (i.e., the 1970s; *Standards for Educational and Psychological Tests* published by the American Psychological Association, American Educational Research Association, and by the National Council on Measurement in Education). Item development was informed by patient interviews and by field testing of items. Even though this questionnaire was developed 30 years ago, the approach conforms to the more recently published *Draft Guidance for Industry, Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims by the FDA.* The BPI has since been validated for use as a brief and meaningful pain assessment tool for noncancer pain conditions as well (17,18).

Acceptability. Acceptability was assessed in a noncancer pain population. The BPI was readily accepted by patients, was not associated with excessive missing data, and did not have problematic floor/ceiling effects (20).

Reliability. Internal consistency for the Pain Severity Score and for the Interference scale has been reported as being 0.85 and 0.88, respectively, in noncancer pain populations (18). Test–retest reliability has been assessed for both cancer and noncancer forms of pain and for over varying time frames. For very short time intervals (e.g., 30–60 minutes), the test–retest reliability was 0.98 for pain severity and 0.97 for pain interference (21). Test–retest reliability for daily administration ranges between 0.83–0.88 for pain severity and between 0.83–0.93 for pain interference (22). FM is considered to be a form of noncancer or musculoskeletal pain and as such these metrics could be applied to FM; however, formal assessment of reliability of the BPI in FM is not available.

Validity. Item analysis has consistently revealed a 2-factor structure (severity or intensity and interference) in more than 36 studies of the BPI across multiple languages for both cancer and noncancer pain populations (23). Construct validity of the BPI has been supported for the generic assessment of pain as well as specifically for low back pain, rheumatoid arthritis (17), and osteoarthritis (20). In a sample of patients with arthritis, the BPI pain severity score correlated (r = 0.74) with the bodily pain scale of the Short Form 36, a generic measure of pain intensity, and (r = 0.77) with the Chronic Pain Grade Intensity scale, another generic pain intensity measure. The BPI Interference scale from this same sample correlated (r = 0.81) with the Chronic Pain Grade disability scale, and (r = 0.69) with the Health Assessment Questionnaire disability index, a disease-specific measure of functional interference (17).

Ability to detect change. The BPI has demonstrated response to change in response to many forms of pharmacologic and nonpharmacologic treatments (23). In chronic pain states, generally, an improvement of 30% or 2–3 points improvement is considered to be a clinically meaningful change (24–26). In a pooled analysis across 12 weeks of treatment from 4 randomized controlled trials of duloxetine for fibromyalgia (FM), the BPI “average pain” and the “Pain Severity Score” was anchored against the Patient Global Impression of Improvement scale (PGI-I). Anchor-based minimum clinically important differences for the “average” pain and for the PGI-I were calculated based upon the difference in mean change from baseline to end point resulting in values of 2.1 and 2.2 points, respectively. This amount of change was associated with 32% and 34% reductions in pain from the baseline scores, respectively (27).

References. The user manual for the BPI contains a reference listing of 72 studies supporting the valid use of the BPI across a wide variety of chronic pain conditions, including FM (23).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BPI was designed to monitor change in pain (and its impact) over time. Numerous studies support its validity to function in this capacity.

Caveats and cautions. The BPI is an industry standard for the generic assessment of both cancer and noncancer pain conditions and contains few flaws in terms of psychometrics, ease of administration, or utility. Far more is known about the psychometrics of the Pain Severity scale and the Pain Interference scale than about the other features of the questionnaire (pain relief, body map, etc.). These other features are often not reported in trials using this instrument. Reports specifically focused upon the psychometric evaluation of the BPI in FM are not available; however, FM is classified as a chronic noncancer musculoskeletal pain condition and the validity of the BPI is supported for the generic assessment of pain intensity and interference.

Clinical usability. The BPI is recommended for use in clinical settings to monitor the severity and impact of pain generically.

Research usability. The BPI is recommended as tool of choice for the assessment of pain in clinical pain trials (28). It is easily administered and has low patient burden.

MULTIDIMENSIONAL FATIGUE INVENTORY (MFI-20)

Description

Purpose. The MFI-20 was introduced 1995 (29) as a measure of fatigue severity. Fatigue is perhaps the most common complaint heard by clinicians. Apart from the everyday use of the term to describe normal tiredness, it can be used to indicate the presence of disease (29). Therefore, the MFI-20 was developed to function as an index of disease, as a diagnostic criterion, or as an outcome variable when a treatment is being evaluated.

Content. The MFI-20 possesses 5-factor analytically confirmed subscales assessing general fatigue, physical fatigue, reduced activity, reduced motivation, and mental fatigue. The MFI differs from other multidimensional fatigue measures by purposely retaining a relatively short list of items, and by eliminating somatic items.

Number of items. The MFI-20 contains 20 items.

Response options/scale. The MFI-20 uses the same response set for each of the 20 items. The respondent is asked to mark an X in 1 of 5 boxes arranged linearly and anchored by “yes, that is true” at one pole to “no, that is not true” at the opposite pole. Scoring of scales requires some items to be reversed such that a higher score on each scale is indicative of greater fatigue.

Recall period for items. The time frame is somewhat nonspecific as the questionnaire queries for symptoms occurring “lately.”

Endorsements/examples of use. The MFI-20 has been used in numerous clinical populations, including cancer (30), Sjögren’s syndrome (31), craniopharyngioma (32), myelodysplastic patients (33), chronic fatigue syndrome

(29), fibromyalgia (FM) (34), and general chronic pain (35). It has also been validated for use in nonclinical samples, including psychology students, medical students, Army recruits, and junior physicians (29).

Practical Application

How to obtain. The MFI-20 is available from the author: E. M. A. Smets, Academic Medical Centre, University of Amsterdam, Department of Medical Psychology, Amsterdam, The Netherlands.

Method of administration. The MFI-20 is a self-report questionnaire.

Scoring. Each scale can be calculated by summing the specific items within each scale. Some items need to be reverse scored prior to summing.

Score interpretation. Each scale contains 4 items with a maximum value of 20 (i.e., each item is endorsed with a "5") and a minimum value of 4 (i.e., each item is endorsed with a "1"). Higher scores on each scale indicate more fatigue severity.

Respondent burden. It takes approximately 5 minutes to complete the MFI-20.

Administrative burden. The MFI-20 is easily administered by handing the questionnaire to the participant. Scoring is accomplished by reverse scoring the required items and then summing each of the 5 scales. Scoring can be completed in <5 minutes.

Translations/adaptations. Validated translations are available for the following languages: English, Swedish, French, and German.

Psychometric Information

Method of development. At the time of development, both 1-dimensional and multidimensional measures of fatigue existed but were quite lengthy and confounded by somatic items. With a consideration of the legacy measures of the time, development of the MFI was initiated by postulating the existence of 5 dimensions of fatigue. Items were generated and then field tested in a diverse group of individuals expected to experience a wide range of fatigue, including individuals with cancer, individuals with chronic fatigue syndrome, first-year medical and psychology students, junior physicians, and Army recruits. Confirmatory factor analysis supported the retention of the 5-dimensional model inherent in this instrument (29).

Acceptability. The MFI is not associated with excessive missing data problems or with or floor/ceiling effects (36).

Reliability. In the original validation study, internal consistency (Cronbach's alpha) for the 5 scales ranged between 0.53–0.93 with the average being 0.80 (29). A more recent validation study of the MFI-20 conducted in the US with a general population sample found the following Cronbach's alpha values: general fatigue (0.83), physical fatigue (0.81), reduced activity (0.82), reduced motivation (0.71), and mental fatigue (0.86) (36). Internal consistency of a total of all 20 items was 0.93. Test–retest reliability has not been reported.

Validity. Confirmatory factor analysis has repeatedly found a 5-factor solution as best fitting the data (i.e., general fatigue, physical fatigue, reduced motivation, reduced activity, mental fatigue), each with adjusted goodness of fit indexes above 0.90 (30). Convergent validity was supported by comparing each scale to a visual analog scale (VAS) assessing fatigue. Associations were all significant with the general fatigue scale having the strongest relationship (30). Construct validity for each scale in association with other relevant constructs has been supported in several validation studies for the MFI-20 (29,30,36).

Ability to detect change. Formally established minimum clinically important differences have not been published for the MFI-20 in FM, however each of the scales appear to be responsive to treatment changes, especially the general fatigue scale (30).

References. There is no specific user manual but the original manuscript provides details on the development and psychometrics of the instrument (29).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The MFI-20 is a brief measure of fatigue that appears to capture relevant dimensions of fatigue severity. It has been used successfully in FM and appears to be a good marker of illness across a broad range of medical illnesses. While not as brief as a single-item VAS (as is commonly used), the MFI-20 correlates well with these measures but offers greater clarification of the type of fatigue being experienced and offers better assessment precision than single-item measures. The MFI does a good job of capturing the experience of fatigue across multiple dimensions without being contaminated by constructs such as functional status (i.e., the functional impact of fatigue), which is better assessed by functional status measures.

Caveats and cautions. Five levels of "yes, that is true" to "no, that is not true" represent a difficult response set for some patients to interpret.

Clinical usability. The MFI-20 may be too lengthy for the typical clinic where a briefer screen may be more appropriate. If however there is a desire to track specific forms of fatigue over time, then this is an appropriate measure.

Research usability. The MFI-20 is recommended for use in clinical trials of interventions targeting fatigue. It has been used successfully in clinical trials of FM (37).

MEDICAL OUTCOMES STUDY (MOS) SLEEP SCALE

Description

Purpose. The MOS Sleep Scale was originally developed as part of the MOS, which was a 4-year observational study of health outcomes for chronically ill patients. The MOS Sleep Scale represents the portion of this larger assessment protocol that specifically focused upon sleep (38). The MOS Sleep Scale is a non–disease-specific measure of multiple aspects of sleep problems.

Content. The MOS Sleep Scale is a 12-item measure assessing 6 domains of sleep: 1) sleep disturbance (e.g., the ability to fall and stay asleep), 2) sleep adequacy (e.g., sleeping enough to feel rested and restored), 3) sleep quantity (e.g., the number of hours slept), 4) somnolence (e.g., daytime sleepiness), 5) snoring, and 6) shortness of breath or headache.

Number of items. The MOS Sleep Scale contains 12 items in its original form; this form has been used in the context of fibromyalgia (FM) clinical trials (37,39) and will be the focus of this review. A briefer 6-item version is also available from the publisher.

Response options/scale. The MOS Sleep Scale uses a variety of response sets. Item 1 queries about how long it takes to fall asleep. Response options are blocked into "0–15 minutes," "16–30 minutes," "31–45 minutes," "46–60 minutes," and "more than 60 minutes." Item 2 queries about how many hours of sleep were obtained on average over the past 4 weeks. This is an open-ended question ranging between 0–24 hours. The remaining 10 items use a 6-point response set based upon the following values and anchors (1 = all of the time, 2 = most of the time, 3 = a good bit of the time, 4 = some of the time, 5 = a little of the time, and 6 = none of the time).

Recall period for items. The time frame for each item is the past 4 weeks. An acute 1-week recall version is also available.

Endorsements/examples of use. The MOS Sleep Scale has been used in numerous nonclinical and clinical populations, including a general US sample (40), cancer pain (41), restless legs syndrome (42), overactive bladder (43), rheumatoid arthritis (44), dialysis (45), neuropathic pain (46), and FM (47).

Practical Application

How to obtain. The MOS Sleep Scale is available from its publisher, Quality Metric. More information can be found at URL: QualityMetric.com. It is recommended that the interested user contact the publisher to learn about potential pricing or licensing agreements associated with the use of this instrument.

Method of administration. The MOS Sleep Scale is a self-report questionnaire.

Scoring and score interpretation. Each scale can be hand scored. Some scales are single items and do not require scoring while others require items to be reversed and summed. Each scale (except sleep quantity) is recalibrated onto a 0–100 scale. For most scales, higher scores indicate worse sleep problems. The exceptions are sleep adequacy and sleep quantity where lower scores indicate worse sleep problems. The MOS Sleep Scale can be aggregated to produce 2 summary indices, the Sleep Problems Index II (9 items) and the Sleep Problems Index I (6 items). Each of these indices integrates the domains of sleep disturbance, sleep adequacy, shortness of breath, and somnolence into a single score. The difference between Sleep Problems Index 1 and 2 is simply length rather than domain coverage; potentially overlapping items were eliminated in Index 1. Higher scores on either index are indicative of worse sleep problems.

Respondent burden. It takes approximately 3–5 minutes to complete the MOS Sleep Scale.

Administrative burden. The MOS Sleep Scale is easily administered by handing the questionnaire to the participant. Scoring requires some reverse scoring, recalibrating scales onto a 0–100 scale, and aggregating the 2 summary indices. It can take 5–7 minutes to score.

Translations/adaptations. The 12-item version is available in 85 languages, which are available from the publisher.

Psychometric Information

Method of development. The MOS Sleep Scale was developed using an extensive review of the published sleep literature resulting in the selection of the domains contained in the scaling of this instrument. The intent was to construct an instrument that would identify sleep problems across sleep-related diseases and associated illnesses rather than being specific to any one type of problem. The scale was initially field tested in a large sample of healthy individuals as well as individuals with a variety of chronic illnesses associated with the MOS (42).

Acceptability. In an evaluation of the MOS Sleep Scale in neuropathic pain, missing items were observed in <10% of the sample. Ceiling and floor effects for each item were acceptable (i.e., <0.50% of all cases). A single item, "awakening short of breath," accounted for much of the problems in scaling properties (46). A second study found similar characteristics for a restless legs syndrome sample with <5% of cases experiencing floor or ceiling effects for the scale as a whole and <20% experiencing floor or ceiling effects for summed scales and <50% for individual items (42).

Reliability. Taken from the neuropathic pain study above (46), Cronbach's alpha ranged between 0.64–0.84 for the MOS sleep subscales. In restless legs syndrome all scales exceeded Cronbach's alpha of 0.70 with the exception of somnolence ($\alpha = 0.66$) (42). In a study of FM all multi-item scales (i.e., sleep disturbance, sleep adequacy, somnolence, and summary indices) exceeded $\alpha = 0.70$ (47).

Validity. Support for construct validity was identified in the restless legs syndrome study where worsening MOS sleep domain scores correlated strongly with worsening indices of quality of life (42). Multitrait scaling was used in the neuropathic pain sample to support convergent and divergent construct validity (46) and recently, confirmatory factor analysis has supported the factorial structure of the MOS Sleep Scale in FM (47). Qualitative interviews (i.e., cognitive debriefing) with patients with FM demonstrated that the MOS Sleep Scale was of relevance to individuals with FM and adequately captured the experience of sleep difficulties arising in FM (48). Additional work associated with criterion validity is needed for the MOS Sleep Scale when specifically applied to FM.

Ability to detect change. In a neuropathic pain sample, the minimal important difference for the 9-item Problem Index 2 was 5.1 (scale 0–100) (46). This is considered a moderate effect (0.65) and corresponds to the corrected change in a group of patients demonstrating change contrasted to the variation observed in a group of patients demonstrating no change. A study in FM reported a clinically important difference (CID) for the sleep disturbance subscale as being 7.9 points (47). CID was calculated by examining differences from baseline as a function (i.e., anchored) of the Patient Global Impression of Change.

References. The publisher, Quality Metric, provides references regarding the development and psychometrics of this instrument.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The MOS Sleep Scale is widely used and is a generic measure of sleep problems that can be used to compare different clinical populations to one another on a common metric. The questionnaire is brief, responsive to change, and has been used in FM.

Caveats and cautions. The items do not use a uniform structure and the scoring is relatively complex given its brevity. The interpretation of the 2 composite indices is not completely obvious except that they are a combination of the assessed domains. Additional data supporting validity and responsiveness to change in FM are desirable.

Clinical usability. The MOS Sleep Scale can be used clinically to monitor changes in sleep across time and within broadly based domains of sleep problems; however, it is a bit lengthy for routine clinical use (48).

Research usability. The MOS Sleep Scale can be used to monitor treatment effects and appears to be sensitive to change both in sleep and in overall quality of life when sleep or other related symptoms improve or worsen.

MULTIPLE ABILITY SELF-REPORT QUESTIONNAIRE (MASQ)

Description

Purpose. The MASQ was purposely designed to assess the self-perception of cognitive difficulties in contrast to the more traditional "objective" neuropsychological assessment by a clinician (49). At the time of development, there were several measures of perceived memory problems, but other relevant areas of cognition lacked a valid self-appraisal tool.

Content. The MASQ contains items about perceived cognitive difficulties in 5 domains of clinical neuropsychological evaluation. The domains of the MASQ along with neuropsychological tests commonly used to index each domain are (50) language (L): Boston Naming Test, Controlled Oral Word Association (C, F, and L words and animals); visual-perceptual ability (VP): Wechsler Adult Intelligence Test Revised (WAIS-R; Block Design, Judgment of Line Orientation); verbal memory (VM): California Verbal Learning Test (Trials 1–5 total, Long Delay Free Recall), Wechsler Memory Scale Revised (WMS-R; Logical Memory I and II); visual-spatial memory (VSM): Rey-Osterrieth Complex Figure (immediate and delayed reproduction), WMS-R Visual Reproduction I and II; and attention/concentration (AC): Stroop Color-Word Test, WAIS-R Arithmetic, WAIS-R Digit Span.

Number of items. The MASQ contains 38 items.

Response options/scale. The MASQ uses the same 5-point response set for all items verbally anchored by "never," "rarely," "sometimes," "usually," and "always." The 5 scales (i.e., L, VP, VM, VSM, AC) are summed. A total score is produced by combining all items.

Recall period for items. No time frame is indicated on the original form.

Endorsements/examples of use. The MASQ has been used to assess perceived cognitive problems in several populations, including the following: epilepsy (49–51), adjuvant chemotherapy for breast cancer (52), breast cancer survivors (53), and fibromyalgia (FM).

Practical Application

How to obtain. The MASQ is available through the instrument's author: Michael Seidenberg, Department of Psychology, UHS/CMS, 3333 Green Bay Road, North Chicago, IL 60064.

Method of administration. The MASQ is administered as a self-report questionnaire.

Scoring. Each item is scaled between 1–5. Nearly half of the items require reverse scoring prior summing. Each scale is then summed. A total score containing all items is also possible. The maximum score for the total score is 190 (i.e., 38 items × 5). Scales containing 8 items (i.e., L, VM, VSM, AC) have a maximum score of 40 and VP (6 items) has a maximum score of 30.

Score interpretation. Higher scores on any scale indicate greater perceived difficulties with that cognitive domain.

Respondent burden. It takes approximately 10 minutes to complete the MASQ.

Administrative burden. The MASQ is easily administered by handing the questionnaire to the participant. Scoring is relatively simple but does require reverse scoring for nearly half of the items before summing.

Translations/adaptations. The MASQ is available in English.

Psychometric Information

Method of development. The initial version of the MASQ contained 48 items based upon clinical experience and a review of published questionnaires at the time of development. Content relevance was evaluated by 8 clinical neuropsychologists and 1 neuropsychiatrist with respect to the cognitive function depicted by each item. Agreement among raters for the retained items supports the content validity of each item.

Acceptability. In the development sample, 22% missed at least 1 item. Ceiling and floor effects were not reported.

Reliability. In the original validation sample, Cronbach's alpha was 0.92 for the total score. Internal consistency was above 0.70 for each of the individual scales (49). In other clinical samples, similar reliability estimates have been reported (e.g., $\alpha = 0.93$ for total and ranged from 0.72–0.79 for subscales in breast cancer survivors) (53). In the original validation study, 2-month test–retest reliability for the entire questionnaire was 0.71 and ranged between 0.55 (L) and 0.74 (VM) (49). Test–retest data and internal consistency data is not available for FM.

Validity. In the original development of the MASQ, items were field tested in 2 samples, individuals with unilateral temporal-lobe epilepsy and healthy normal individuals. Support for concurrent validity came from higher MASQ scores being associated with poorer performance on neuropsychological tests in both samples but with greater perceived difficulties being observed in the clinical sample. These studies support the idea that perceived cognitive difficulties correspond to more objectively assessed indices of the same constructs (49). In a study comparing individuals with FM to healthy controls, individuals with FM scored significantly higher on each MASQ subscale than did healthy controls (54). Studies assessing the criterion validity of the MASQ with objective neuropsychological performance tests in FM are not currently available.

Ability to detect change. Reliable change indices and standard regression-based change norms have been established for the MASQ for use in cases of epilepsy (51). The MASQ has also demonstrated response to change in clinical trials of therapeutics for FM (e.g., milnaciparan) (55).

References. Original support for the MASQ is found in the work by Seidenberg et al (49).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. Fibro fog is a common complaint among individuals with FM. Often only the memory aspects are assessed, but patients complain of broader deficits that are covered by the MASQ. The MASQ can be useful in tracking the varied manifestations of dyscognition in FM that are related to the different symptoms that characterize FM.

Caveats and cautions. The length of this instrument at 38 items may be prohibitive in settings where multiple domains of clinical relevance need to be efficiently measured. The MASQ has not been as rigorously developed or tested as the other measures reviewed in this article, but is one of the few measures currently available to assess this important aspect of FM.

Clinical usability. The MASQ appears to capture multiple aspects of fibro fog. Patients express a desire to have this domain assessed; yet, there are few instruments aside from the MASQ that are available for this purpose.

Research usability. The MASQ has been used in several large scale clinical trials of therapeutics for FM supporting is ability to assess and detect change in perceived cognitive difficulties.

AUTHOR CONTRIBUTIONS

Both authors were involved in drafting the article or revising it critically for important intellectual content, and both authors approved the final version to be published.

REFERENCES

1. Wolfe F, Clauw DJ, Fitzcharles MA, Goldenberg DL, Katz RS, Mease P, et al. The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. Arthritis Care Res (Hoboken) 2010;62:600–10.
2. Arnold LM, Crofford LJ, Mease PJ, Burgess SM, Palmer SC, Abetz L, et al. Patient perspectives on the impact of fibromyalgia. Patient Educ Couns 2008;73:114–20.
3. Mease PJ, Arnold LM, Crofford LJ, Williams DA, Russell IJ, Humphrey L, et al. Identifying the clinical domains of fibromyalgia: contributions from clinician and patient Delphi exercises. Arthritis Rheum 2008;59: 952–60.
4. Arnold LM, Mease PJ, Williams DA, Martin SA, Wang F, Emir B, et al. Development of responder definitions for fibromyalgia clinical trials. Arthritis Rheum 2010;62:S38.
5. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63: 1179–94.
6. Bennett RM. The Fibromyalgia Impact Questionnaire (FIQ): a review of its development, current version, operating characteristics and uses. Clin Exp Rheumatol 2005;23:S154–62.
7. Burckhardt CS, Clark SR, Bennett RM. The Fibromyalgia Impact Questionnaire: development and validation. J Rheumatol 1991;18:728–33.
8. Bennett RM, Friend R, Jones KD, Ward R, Han BK, Ross RL. The Revised Fibromyalgia Impact Questionnaire (FIQR): validation and psychometric properties. Arthritis Res Ther 2009;11:R120.
9. Wolfe F, Hawley DJ, Goldenberg DL, Russell IJ, Buskila D, Neumann L. The assessment of functional impairment in fibromyalgia (FM): Rasch analyses of 5 functional scales and the development of the FM Health Assessment Questionnaire. J Rheumatol 2000;27:1989–99.
10. Arnold LM, Rosen A, Pritchett YL, D'Souza DN, Goldstein DJ, Iyengar S, et al. A randomized, double-blind, placebo-controlled trial of duloxetine in the treatment of women with fibromyalgia with or without major depressive disorder. Pain 2005;119:5–15.
11. Arnold LM, Russell IJ, Diri EW, Duan WR, Young JP Jr, Sharma U. A 14-week, randomized, double-blinded, placebo-controlled monotherapy trial of pregabalin in patients with fibromyalgia. J Pain 2008; 9:792–805.
12. Arnold LM, Gendreau RM, Palmer RH, Gendreau JF, Wang Y. Efficacy and safety of milnacipran 100 mg/day in patients with fibromyalgia: results of a randomized, double-blind, placebo-controlled trial. Arthritis Rheum 2010;62:2745–56.
13. Arnold LM, Clauw DJ, Wohlreich MM, Wang F, Ahl J, Gaynor PJ, et al. Efficacy of duloxetine in patients with fibromyalgia: pooled analysis of 4 placebo-controlled clinical trials. Prim Care Companion J Clin Psychiatry 2009;11:237–44.
14. Bennett RM, Bushmakin AG, Cappelleri JC, Zlateva G, Sadosky AB. Minimal clinically important difference in the fibromyalgia impact questionnaire. J Rheumatol 2009;36:1304–11.
15. Cleeland CS. Measurement and prevalence of pain in cancer. Semin Oncol Nurs 1985;1:87–92.
16. Daut RL, Cleeland CS, Flanery RC. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain 1983;17:197–210.
17. Keller S, Bann CM, Dodd SL, Schein J, Mendoza TR, Cleeland CS. Validity of the brief pain inventory for use in documenting the outcomes of patients with noncancer pain. Clin J Pain 2004;20:309–18.
18. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the Brief Pain Inventory for chronic nonmalignant pain. J Pain 2004;5:133–7.
19. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9–19.
20. Williams VS, Smith MY, Fehnel SE. The validity and utility of the BPI interference measures for evaluating the impact of osteoarthritic pain. J Pain Symptom Manage 2006;31:48–57.
21. Radbruch L, Loick G, Kiencke P, Lindena G, Sabatowski R, Grond S, et al. Validation of the German version of the Brief Pain Inventory. J Pain Symptom Manage 1999;18:180–7.
22. Mendoza T, Mayne T, Rublee D, Cleeland C. Reliability and validity of a modified Brief Pain Inventory short form in patients with osteoarthritis. Eur J Pain 2006;10:353–61.
23. Cleeland C. The Brief Pain Inventory: user guide. Houston: MD Anderson Cancer Center; 2009.
24. Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL. Defining the clinically important difference in pain outcome measures. Pain 2000; 88:287–94.
25. Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001;94:149–58.
26. Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, et al. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain 2003;106:337–45.
27. Mease PJ, Spaeth M, Clauw DJ, Arnold LM, Bradley LA, Russell IJ, et al.

Estimation of minimum clinically important difference for pain in fibromyalgia. Arthritis Care Res (Hoboken) 2011;63:821–6.

28. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 2008; 9:105–21.
29. Smets EM, Garssen B, Bonke B, De Haes JC. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39:315–25.
30. Smets EM, Garssen B, Cull A, de Haes JC. Application of the multidimensional fatigue inventory (MFI-20) in cancer patients receiving radiotherapy. Br J Cancer 1996;73:241–5.
31. Barendregt PJ, Visser MR, Smets EM, Tulen JH, van den Meiracker AH, Boomsma F, et al. Fatigue in primary Sjogren's syndrome. Ann Rheum Dis 1998;57:291–5.
32. Dekkers OM, Biermasz NR, Smit JW, Groot LE, Roelfsema F, Romijn JA, et al. Quality of life in treated adult craniopharyngioma patients. Eur J Endocrinol 2006;154:483–9.
33. Jansen AJ, Essink-Bot ML, Beckers EA, Hop WC, Schipperus MR, Van Rhenen DJ. Quality of life measurement in patients with transfusion-dependent myelodysplastic syndromes. Br J Haematol 2003;121:270–4.
34. Ericsson A, Mannerkorpi K. Assessment of fatigue in patients with fibromyalgia and chronic widespread pain: reliability and validity of the Swedish version of the MFI-20. Disabil Rehabil 2007;29:1665–70.
35. Fishbain DA, Lewis J, Cole B, Cutler B, Smets E, Rosomoff H, et al. Multidisciplinary pain facility treatment outcome for pain-associated fatigue. Pain Med 2005;6:299–304.
36. Lin JM, Brimmer DJ, Maloney EM, Nyarko E, Belue R, Reeves WC. Further validation of the Multidimensional Fatigue Inventory in a US adult population sample. Popul Health Metr 2009;7:18.
37. Clauw DJ, Mease P, Palmer RH, Gendreau RM, Wang Y. Milnacipran for the treatment of fibromyalgia in adults: a 15-week, multicenter, randomized, double-blind, placebo-controlled, multiple-dose clinical trial. Clin Ther 2008;30:1988–2004.
38. Hays RD, Stewart A. Sleep measures. In: Stewart A, Ware J, editors. Measuring functioning and well-being: the medical outcomes study approach. Durham (NC): Duke University Press; 1992. p. 235–59.
39. Mease PJ, Clauw DJ, Gendreau RM, Rao SG, Kranzler J, Chen W, et al. The efficacy and safety of milnacipran for treatment of fibromyalgia. a randomized, double-blind, placebo-controlled trial. J Rheumatol 2009; 36:398–409.
40. Hays RD, Martin SA, Sesti AM, Spritzer KL. Psychometric properties of the Medical Outcomes Study Sleep measure. Sleep Med 2005;6:41–4.
41. Payne R, Mathias SD, Pasta DJ, Wanke LA, Williams R, Mahmoud R. Quality of life and cancer pain: satisfaction and side effects with transdermal fentanyl versus oral morphine. J Clin Oncol 1998;16: 1588–93.
42. Allen RP, Kosinski M, Hill-Zabala CE, Calloway MO. Psychometric evaluation and tests of validity of the Medical Outcomes Study 12-item Sleep Scale (MOS sleep). Sleep Med 2009;10:531–9.
43. Coyne KS, Zhou Z, Bhattacharyya SK, Thompson CL, Dhawan R, Versi E. The prevalence of nocturia and its effect on health-related quality of life and sleep in a community sample in the USA. BJU Int 2003;92: 948–54.
44. Wolfe F, Michaud K, Li T. Sleep disturbance in patients with rheumatoid arthritis: evaluation by medical outcomes study and visual analog sleep scales. J Rheumatol 2006;33:1942–51.
45. Unruh ML, Hartunian MG, Chapman MM, Jaber BL. Sleep quality and clinical correlates in patients on maintenance dialysis. Clin Nephrol 2003;59:280–8.
46. Rejas J, Ribera MV, Ruiz M, Masrramon X. Psychometric properties of the MOS (Medical Outcomes Study) Sleep Scale in patients with neuropathic pain. Eur J Pain 2007;11:329–40.
47. Cappelleri JC, Bushmakin AG, McDermott AM, Dukes E, Sadosky A, Petrie CD, et al. Measurement properties of the Medical Outcomes Study Sleep Scale in patients with fibromyalgia. Sleep Med 2009;10: 766–70.
48. Martin S, Chandran A, Zografos L, Zlateva G. Evaluation of the impact of fibromyalgia on patients' sleep and the content validity of two sleep scales. Health Qual Life Outcomes 2009;7:64.
49. Seidenberg M, Haltiner A, Taylor MA, Hermann BB, Wyler A. Development and validation of a Multiple Ability Self-Report Questionnaire. J Clin Exp Neuropsychol 1994;16:93–104.
50. Banos JH, LaGory J, Sawrie S, Faught E, Knowlton R, Prasad A, et al. Self-report of cognitive abilities in temporal lobe epilepsy: cognitive, psychosocial, and emotional factors. Epilepsy Behav 2004;5:575–9.
51. Martin R, Griffith HR, Sawrie S, Knowlton R, Faught E. Determining empirically based self-reported cognitive change: development of reliable change indices and standardized regression-based change norms for the multiple abilities self-report questionnaire in an epilepsy sample. Epilepsy Behav 2006;8:239–45.
52. Donovan KA, Small BJ, Andrykowski MA, Schmitt FA, Munster P, Jacobsen PB. Cognitive functioning after adjuvant chemotherapy and/or radiotherapy for early-stage breast carcinoma. Cancer 2005;104: 2499–507.
53. Jim HS, Donovan KA, Small BJ, Andrykowski MA, Munster PN, Jacobsen PB. Cognitive functioning in breast cancer survivors: a controlled comparison. Cancer 2009;115:1776–83.
54. Williams DA, Clauw DJ, Glass JM. Perceived cognitive dysfunction in fibromyalgia syndrome. J Musculoskelet Pain 2011;19:66–75.
55. Owen RT. Milnacipran hydrochloride: its efficacy, safety and tolerability profile in fibromyalgia syndrome. Drugs Today (Barc) 2008;44: 653–60.

Summary Table for Fibromyalgia Measures

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
Fibromyalgia Impact Questionnaire	Disease-specific assessment of multiple facets of fibromyalgia: physical function, overall impact, and symptoms	Patient self-report	Completion in 3–5 minutes	Hand scored in under 5 minutes	Range: 0–10, higher scores indicate greater impairment; total score range: 0–100	Internal consistency: >0.90; test–retest: 0.56–0.95	Factor structure confirmed; construct validity supported	Minimum clinically important difference: 14% change on total score	Only disease-specific measure in fibromyalgia; useful in both clinical practice and clinical research; covers multiple aspects of fibromyalgia	Functional scale biased to high levels of disability; newer Fibromyalgia Impact Questionnaire Revised not tested in clinical trials
Brief Pain Inventory	General assessment of pain: severity interference, medications, relief from medications, presence of pain, and pain distribution	Patient self-report or interview	Completion in 5 minutes	Hand scored by summing in under 5 minutes	Pain severity (range 0–10) with higher scores indicating greater pain, interference (range 0–10) with higher scores indicating greater interference	Internal consistency: 0.85–0.88 for severity and interference, respectively; test–retest: 0.83–0.98	Factor structure confirmed; construct validity supported	Minimum clinically important difference: 30% change on severity score	Rigorous development; widely used clinically; widely used in clinical trials, recommended by Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials group	Scores other than severity and interference often not reported
Multidimensional Fatigue Inventory	Profound fatigue associated with illness: general, physical, mental, reduced motivation, and reduced activity	Patient self-report	Completion in 5 minutes	Hand scored in 5 minutes	Each scale contains a minimum value of 4 and a maximum of 20; higher scores indicate greater fatigue	Internal consistency: average is 0.80 for scales; total score internal consistency = 0.93	Factor structure confirmed; construct validity supported	No minimum clinically important difference	Used in fibromyalgia clinical trials; good metric properties for clinical and research use	Response set can be confusing to respondents
Medical Outcomes Study Sleep Scale	A generic assessment of sleep problems: disturbance, duration, adequacy, somnolence, snoring, shortness of breath, and summary indices	Patient self-report	Completion in 3–5 minutes	Hand scored in 5–7 minutes	Most scales range 0–100, scales are mixed with regard to the interpretation of scale values	Internal consistency range 0.64–0.84; in fibromyalgia, internal consistency = 0.70	Factor structure confirmed; construct validity supported	Minimal important difference 5.1	Generic instrument; widely used in clinical research; responsive to change; can be used in clinical settings	Scoring and meaning are not intuitive
Short Form 36 physical and mental component scores	Physical and mental health functional status	Self-report, interview, or online	Completion in 5–7 minutes	Scored by publisher, software online, or distributed	Lower scores indicate worse functional status; scores are norm-based on US population with a mean of 50; a 10-point change is equivalent to 1 SD	Internal consistency >0.90; test– retest: r ranges 0.60–0.80 for 2 weeks	Factor structure confirmed; construct validity supported	Minimum clinically important difference in fibromyalgia: 6-point change	Extremely well-developed instrument; useful in clinical work; useful in clinical trials	Can be costly to use; scoring is complex and requires software
Multiple Ability Self-Report Questionnaire	Perceived cognitive deficits: language, visual-perceptual, verbal memory, visual-spatial memory, and attention-concentration	Self-report	Completion in 10 minutes	Hand scored involving item reversal and summing	Higher scores indicate more dysfunction; 8-item scales have a max value of 40; 6-item scales have max value of 30; total score has max value of 190	Internal consistency >0.70, test–retest: 0.72–0.79	Construct validity supported using convergent and divergent methods	Minimum clinically important differences have not been established for fibromyalgia	Only instrument that assesses multiple aspects of dyscognition in fibromyalgia, useful clinically; useful for clinical trials of fibromyalgia	Somewhat lengthy
Hospital Anxiety and Depression Scale	Anxiety and depression screener for nonpsychiatric populations	Self-report	Completion in 2–5 minutes	Hand scoring by summing items (1–2 minutes)	Higher scores indicate more anxiety and depressive symptoms; 0–7 normal, 8–10 mild, >11 probable caseness	Internal consistency 0.80–0.93	Factor structure confirmed; construct validity supported	Minimum clinically important difference not established for fibromyalgia	Avoids somatic symptoms and extreme psychiatric symptoms more appropriate for a nonpsychiatric population; quick to administer and score in the clinic or research setting; has been used in multiple fibromyalgia clinical trials	Does not provide a diagnosis, only an estimate of potential caseness

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S413–S419
DOI 10.1002/acr.20636
© 2011, American College of Rheumatology

Health-Related Quality of Life Measurement in Adult Systemic Lupus Erythematosus

Lupus Quality of Life (LupusQoL), Systemic Lupus Erythematosus-Specific Quality of Life Questionnaire (SLEQOL), and Systemic Lupus Erythematosus Quality of Life Questionnaire (L-QoL)

JINOOS YAZDANY

INTRODUCTION

Throughout the course of their disease, individuals with systemic lupus erythematosus (SLE) face considerable physical, psychological, and social challenges. The disease has profound effects on health-related quality of life (HRQOL), which have been documented extensively in the literature (1). Capturing decrements and improvements in HRQOL has therefore become important in clinical research in SLE, and is advocated by both the US Food and Drug Administration in providing guidance to SLE clinical trialists as well as the Outcome Measures in Rheumatology Clinical Trials group (2,3). Three measures designed to ascertain HRQOL in SLE will be reviewed, the Lupus Quality of Life, Systemic Lupus Erythematosus-Specific Quality of Life questionnaire, and Systemic Lupus Erythematosus Quality of Life questionnaire (L-QoL). These measures were chosen because they were developed and specifically designed as patient-reported outcome measures to assess quality of life in SLE and have all had some published validation testing to date.

Most studies examining HRQOL in SLE have employed generic measures, such as the Medical Outcomes Study Short Form (SF-36) (4). An advantage of generic instruments is that they allow comparison of the HRQOL in SLE to other related conditions or to population norms, something that has been useful in documenting that SLE has similar or worse HRQOL decrements compared to other severe chronic conditions (5). In addition, many generic instruments have undergone extensive validation testing and are adapted in multiple languages and cultures.

However, a disadvantage of employing generic instruments alone in SLE is that they may not adequately capture symptoms or issues that are specific to the disease. This may reduce their sensitivity to detect meaningful changes over time. For example, some, but not all, studies suggest that the SF-36 is insufficiently responsive in longitudinal studies or trials in SLE (6,7), and may lack domains that are particularly relevant to a population with SLE, such as fatigue or sleep (8). The 3 SLE-specific instruments reviewed here have been developed to address some of these potential limitations. As discussed below, preliminary validation work is available for each of these instruments in defined populations.

LUPUS QUALITY OF LIFE (LUPUSQOL)

Description

Purpose. To measure disease specific health-related quality of life (HRQOL) in adult systemic lupus erythematosus (SLE). The original development and validation study was performed in the UK and published by McElhone et al in 2007 (9).

Content. Eight domains are covered, including physical health, emotional health, body image, pain, planning, fatigue, intimate relationships, and burden to others.

Number of items. 34 items total. Individual subscales include physical health (8 items), emotional health (6 items), body image (5 items), pain (3 items), planning (3 items), fatigue (4 items), intimate relationships (2 items), and burden to others (3 items).

Response options/scale. Questionnaire has a 5-point Likert scale response format (0 = all the time, 1 = most of the time, 2 = a good bit of the time, 3 = occasionally, and 4 = never).

Recall period for items. Prior 4 weeks.

Endorsements. None.

Examples of use. The LupusQoL has been used for research purposes in clinical cohorts in the UK and US (10,11). It has not yet been used in a clinical trial in SLE.

Supported by the Arthritis Foundation.

Jinoos Yazdany, MD, MPH: University of California, San Francisco.

Address correspondence to Jinoos Yazdany, MD, MPH, Division of Rheumatology, University of California, San Francisco, Box 0920, San Francisco, CA 94143-0920. E-mail: jinoos.yazdany@ucsf.edu.

Submitted for publication February 9, 2011; accepted in revised form May 23, 2011.

Table 1. Median (IQR [interquartile range]) scores for the 8 Lupus Quality of Life domains in clinic-based samples from the UK (10) and US (11)*

Domain	UK (n = 322)	US (n = 185)
Physical health	65.6 (40.6–81.3)	44.4 (46.4–30.2)
Pain	75.0 (41.7–83.3)	42.9 (50.0–33.3)
Planning	75.0 (50.0–91.7)	48.9 (50.0–41.6)
Intimate relationships	75.0 (37.5–87.5)	53.9 (62.5–37.5)
Burden to others	66.7 (41.7–83.3)	44.5 (50.0–34.3)
Emotional health	75.0 (62.5–87.5)	51.3 (56.2–29.1)
Body image	80.0 (55.0–95.0)	54.2 (56.2–33.3)
Fatigue	56.3 (32.3–68.8)	38.3 (41.6–31.2)

* Values are the median (IQR).

The UK sample was predominantly white and had less severe disease, while the US sample was predominantly African American and had more severe disease. Median domain values for the LupusQoL in these 2 cohorts are shown in Table 1.

Practical Application

How to obtain. Available in the online issue of reference, which is available at http://onlinelibrary.wiley.com/doi/10.1002/art.22881/suppinfo. A web site has also been launched with information regarding obtaining permissions to use the instrument, instructions for scoring and other useful information (www.lupusqol.com).

Method of administration. Written and electronic versions of the questionnaire are available.

Scoring. The mean raw domain score is transformed to scores ranging from 0 (worst HRQOL) to 100 (best HRQOL) by dividing by 4 and then multiplying by 100. The result represents the transformed score for that domain. The authors suggest that transformed domain scores are obtainable when at least 50% of the items are answered. The mean raw domain score is then calculated by totaling the item response scores of the answered items and dividing by the number of answered items. A nonapplicable response is treated as unanswered and the domain score is calculated as indicated above.

Score interpretation. 0 (worst HRQOL) to 100 (best HRQOL).

Respondent burden. Time to complete is <10 minutes. No information on required reading level is provided (the educational attainment of the UK validation cohort was mean ± SD 13.8 ± 3.1 years).

Administrative burden. Time to score is <5 minutes.

Translations/adaptations. A Spanish language version has been adapted and validated (12). A version adapted and validated for a US population is also available (13). Translations into 77 languages from 51 countries are available (see www.lupusqol.com), although these translations do not yet have published psychometric information.

Psychometric Information

Method of development. The original measure was developed and validated by using a mixed qualitative and quantitative approach. Briefly, 30 individuals with SLE participated in semistructured interviews and a combination of thematic analysis from these interviews as well as expert panel feedback was used to generate items. Feedback was sought again from a group of 20 patients to revise draft items. Subscales were generated using principal component analysis. A written survey (either mailed or administered in the clinic) was then used to assess validity and reliability.

It is important to note that the US validation study found a different factor structure for the LupusQoL, with only 5 of the 8 factors having eigenvalues >1 in the analysis (13); eigenvalues are used to measure how much of the variance each successive factor extracts, and only values >1 are generally retained in analyses (14).

Acceptability. Information on readability was not provided, but item response rates were very high (<2% of domains were not scored because of missing responses). However, it is important to note that some domains (i.e., intimate relationships) were not applicable to all respondents (7.3% missing). Floor and ceiling effects were reported for each domain and were reasonable; for all domains except intimate relationships, the percentage of individuals with a score of 0 was <10% (range 2.2–8.6%), and the percentage of individuals with a maximum score of 100 was <30% (range 6.2–28.2%).

Reliability. Individual domains demonstrated good internal consistency (Cronbach's α range 0.88–0.96) in the original validation study as well as in the US and Spanish adaptations. Test–retest reliability of the original LupusQoL was evaluated in a subset of 83 respondents and was good with intraclass correlation coefficients between 0.72–0.93 for the individual domains.

Validity. Concurrent validity was assessed by comparing domain scores of the LupusQoL with other comparable domains of the Medical Outcomes Study Short Form (SF-36), with good correlation (r = 0.71–0.79). Similar results were obtained in the US and Spanish validation studies. Several recent followup studies performed in the UK, US, and Spain demonstrated that the LupusQoL has discriminant validity in that it functions relatively independently as an outcome measure in SLE. These studies found weak or no associations with factors such as disease duration, disease activity and damage (10–12). To assess construct validity, the developers examined LupusQoL scores in relation to disease activity (as measured by the British Isles Lupus Assessment Group) and damage (Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index [SDI]) (9). Patients with more active disease generally reported poorer HRQOL across all domains except fatigue, although the relationship with damage, as measured by the SDI was less clear.

Ability to detect change. Sensitivity to change (responsiveness) and minimum clinically important difference are not yet available, but are subjects of an ongoing study.

Critical Appraisal of Overall Value to the Rheumatology Community

Of the available instruments to assess HRQOL, the LupusQoL has undergone the most extensive validation process and has been modified to be culturally appropriate for the US and Spanish populations. Translations are available in numerous languages, although psychometric evaluations of these translations have not yet been published. The importance of performing such evaluations is evidenced by the differences noted in the UK and US validation studies of the LupusQoL, including the different factor structures identified. The reasons for these differences remain unclear, and further studies are needed to assess the optimal factor structure of the instrument.

Currently, the measure would be most appropriate for cross-sectional evaluations of HRQOL in SLE in the populations in which the measure is validated. Future studies examining the responsiveness of the LupusQoL will elucidate its role in treatment studies of SLE. For longitudinal assessments in observational studies, information about additional psychometric properties, such as response shift bias, may also be useful.

SYSTEMIC LUPUS ERYTHEMATOSUS-SPECIFIC QUALITY OF LIFE QUESTIONNAIRE (SLEQOL)

Description

Purpose. To assess quality-of-life (QOL) in individuals with systemic lupus erythematosus (SLE). The original development and validation study of the English language survey took place in Singapore by Leong et al (6).

Content. Six domains including physical functioning, activities, symptoms, treatment, mood, and self-image.

Number of items. 40 items, including physical functioning (6 items), activities (9 items), symptoms (8 items), treatment (4 items), mood (4 items), and self-image (9 items).

Response options/scale. 7-point response scale (subsections have different anchors, including "not difficult at all" to "extremely difficult," "not at all" to "extremely troubled," and "not at all" to "extremely often").

Recall period for items. 1 week.

Endorsements. None.

Examples of use. The instrument has been used in cross-sectional analyses in SLE clinical cohorts (15,16). In the Brazilian cohort, the mean score was 116 (16).

Practical Application

How to obtain. Contact the author of the original article (6) (Khai_pang_leong@ttsh.com.sg) for the original version or Kok_Ooi_Kong@ttsh.com.sg for the Chinese adaptation (SLEQOL-C) (17).

Method of administration. Written questionnaire.

Scoring. A summary score is derived from the sum of all responses across the domains; alternatively the authors suggest that a summary score can be obtained by taking the mean of each of the 6 subsections. Item weighting is not available and needs to be addressed in future studies given that the current scoring system places greater emphasis on domains with a greater number of items. No specific instruction for dealing with missing values is provided.

Score interpretation. Scores range from 40–280, with higher values corresponding to worse QOL.

Respondent burden. <5 minutes for both the SLEQOL and SLEQOL-C.

Administrative burden. Time to score is not reported.

Translations/adaptations. A Chinese language version (SLEQOL-C) was derived by translation and back-translation, and content validity was examined through interviews with 7 bilingual patients with SLE in Singapore. The study did not demonstrate differential item functioning in the responses of English and Chinese-speaking patients, suggesting successful translation into Chinese (17). Psychometric testing of the SLEQOL-C is not yet available. The SLEQOL has also been culturally adapted and undergone preliminary validation testing in Brazilian-Portuguese using a clinical cohort of 107 patients (16). Interobserver and intraobserver reliability for the adaptation was found to be high, and the measure had good internal consistency. The measure correlated well with the Medical Outcomes Study Short Form (SF-36), suggesting construct validity, and poorly with lupus disease activity and damage measures, suggesting discriminant validity.

Psychometric Information

Method of development. An unspecified number of rheumatologists and nurse clinicians familiar with SLE management generated an initial list of items. Feedback was elicited from 100 patients on these draft items; however, patients were not involved in generation of the items originally. Factor analysis and Rasch model analyses were used to compose the final questionnaire and create subscales. Psychometric properties were tested using responses obtained during routine clinical visits in 275 patients. The characteristics of this clinical cohort included a disease duration of approximately 9 years, a mean ± SD Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) score of 2.7 ± 4.8 and mean ± SD Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index (SDI) score of 0.67 ± 1.1. Patients were from Singapore and English-speaking. A subset of patients had repeat data collection to allow investigation of test–retest reliability and responsiveness.

Acceptability. A minority of participants in the original SLEQOL validation study had low educational attainment (10.5% had no formal education or a primary education only); this number was significantly higher for the SLEQOL-C (44.7% of the sample had no formal education or a primary education only). However, no specific information on readability was provided in the Singapore studies. Research assistants ensured that patients completed items so no missing responses were reported.

An analysis of floor and ceiling effects revealed that the SLEQOL had significant floor effects (good perceived QOL), with 3 of the subsections having between 39 and 44% of individuals reporting good perceived QOL. Ceiling effects were not observed. The SF-36 in the same sample had fewer floor effects, but more significant ceiling effects; for 4 domains, between 28–59% of respondents reported poor QOL.

Reliability. Internal consistency was good (Cronbach's alpha = 0.95 for the summary score, and ranged from 0.76–0.93 for specific subsections).

Test–retest reliability was assessed in 51 patients who repeated the instrument at a 2-week interval. The intraclass correlation coefficient was 0.83 for the summary score, indicating good reliability. However, 4 of the 6 individual domains had intraclass correlation coefficients of <0.6, which indicates only moderate reliability. Reliability in the Brazilian-Portuguese culturally adapted version was high (intraobserver correlation coefficient 0.97 and interobserver correlation coefficient 0.99) (16).

Validity. Although items were generated entirely by health professionals, patient feedback was solicited to add and modify items to assess content validity (6,18,19). Construct validity was investigated by comparing scores on the SLEQOL to the SF-36, Rheumatology Attitudes Index and its helplessness subscale, commonly used physician-assessed disease activity (SLEDAI and Systemic Lupus Activity Measure [SLAM]), and damage indices (SDI). Absent or very weak correlations were demonstrated for the summary score for most SF-36 domains (the strongest correlation being between the SLEQOL physical functioning domain and the SF-36 physical functioning domain at 0.234), suggesting relatively low concurrent validity. Correlations were also weak or absent with the SLAM, SLEDAI, and SDI. However, these data provide evidence of discriminant validity, as the SLEQOL appears to be capturing constructs that are independent of traditional disease activity and damage measures.

Construct validity was supported by an analysis demonstrating that the SLEQOL summary score varied appropriately with self-perceived changes in global QOL.

Ability to detect change. Responsiveness was assessed in a subset of 95 patients who had return clinical visits within a 3-month window. Participants were asked to rate the global change in QOL using a scale anchored from −7 to 7 (where −7 represents "a very great deal worse" and 7 represents "a very great deal better"). Few participants reported significant QOL deterioration, and therefore this group was not analyzed (n = 12). Among individuals who reported QOL improvements or reported no change, responsiveness was assessed using multiple techniques, including the standardized response mean, effective size, Guyatt's coefficient, and relative efficacy. All methods yielded similar results, with the SLEQOL demonstrating greater responsiveness than the individual domains of the SF-36. However, the SLEQOL also demonstrated greater variation of scores in participants who reported unchanged QOL compared to the SF-36, indicating decreased specificity.

Minimum clinically important difference (MCID) was derived using a distributional approach in which SLEQOL scores were anchored to the patient global ratings of changes in their QOL. By taking the mean of the absolute difference of SLEQOL scores in the group of patients who rated their global QOL change as +2 to +3 ("moderately worse" or "a little worse") and −2 to −3 ("moderately better" or "a little better"), the MCID was calculated at approximately 25.

Critical Appraisal of Overall Value to the Rheumatology Community

The strengths of the SLEQOL, which primarily assesses HRQOL, include the fact that information is available on its responsiveness and the MCID. The instrument has good discriminant validity as it appears to function independently from commonly used measures of disease activity, damage, and disease-related attitudes.

Additional studies will be required to further assess and confirm psychometric properties. Psychometric testing of the Chinese language version (SLEQOL-C) is not available. Reliability for the individual domains was only moderate in the original validation study, which suggests that these scores should be used with caution given possible instability. Concurrent validity with the SF-36 is relatively poor, suggesting that the instrument should be used primarily in conjunction with other validated measures of HRQOL. In addition, floor effects should be considered, and as the developers note, the instrument may best be used with a companion generic instrument that does not have substantial floor effects.

SYSTEMIC LUPUS ERYTHEMATOSUS (SLE) QUALITY OF LIFE QUESTIONNAIRE (L-QOL)

Description

Purpose. To provide a needs-based assessment of quality of life (QOL) in systemic lupus erythematosus (SLE). The L-QoL was developed by Doward et al in 2008 (20).

Content. The questionnaire is based on the needs-based QOL model, which posits that life gains its quality from the ability and capacity of individuals to satisfy their needs. Items assess the overall effect of SLE and its treatment on QOL.

Number of items. 25 items in the scale, including items assessing self-care, fatigue, and emotional reactions.

Response options/scale. Dichotomous true/not true response format.

Recall period for items. Not reported.

Endorsements. None.

Examples of use. The instrument has not yet been used in published clinical or observational studies of SLE. The mean ± SD value for the L-QoL in the original validation study performed in the UK was 6.7 ± 6.1.

Practical Application

How to obtain. The instrument is available from the University of Leeds; registration is required. Further information is provided on University of Leeds Psychometric laboratory web site (http://www.leeds.ac.uk/medicine/rehabmed/psychometric/Scales3.htm).

Method of administration. Written questionnaire.

Scoring. Count of symptoms. Higher score on the L-QoL indicates worse QOL. There are no specific instructions for dealing with missing values.

Score interpretation. Score range is 0–25, with higher scores indicating worse QOL.

Respondent burden. <5 minutes.

Administrative burden. Time to score is not reported.

Translations/adaptations. Published adaptations are not available.

Psychometric Information

Method of development. The L-QoL was developed through a multistep process that started with the use of qualitative interviews with 50 individuals with SLE in the UK. Analysis of this qualitative data was used to construct items that were relevant to the needs model, and applicable to all potential respondents. Draft items were revised based on feedback elicited during cognitive interviews with 16 patients. Scaling and psychometric properties were then tested through the use of 2 postal surveys (n = 95 and n = 93). Rasch analysis was conducted to confirm unidimensionality and the absence of differential item functioning.

Acceptability. The readability of the survey is not reported, nor is the educational attainment of the development and validation samples. Overall response rate for the first postal survey was 76%. Missing data were encountered in 14 of 95 (14.7%) responses, although the number of missing items per respondent was relatively low (mean ± SD 2.9 ± 2.7). The presence or absence of floor or ceiling effects was not explicitly analyzed; although the authors provide the range of scores obtained (0–22), the mean ± SD (6.7 ± 6.1), and the median (5.0 ± interquartile range 1.0–11.0).

Reliability. Test–retest reliability was assessed by postal surveys administered 2 weeks apart. The interclass correlation coefficient was 0.95, indicating excellent reliability. Internal consistency using Person-separation reliability was 0.91–0.92.

Validity. Items were derived from patient interviews and were largely phrased in the patients' own words to maximize content validity. Construct validity was demonstrated through examining the relationship between the L-QoL and other measures of disease activity and severity; those with higher perceived disease activity (rated as perceived current disease flare yes/no), higher perceived disease severity (rated on a scale mild/moderate/quite severe), and fair/poor ratings of their general health, had statistically significantly L-QoL scores. Individuals who were unemployed also had lower L-QoL scores, and this reached statistical significance in the second postal sample (but not in the first). In addition, moderate correlations were observed between the L-QoL and Nottingham Health Profile scores (between 0.48 and 0.80).

A Rasch analysis was performed to determine unidimensionality of the scale. This method builds a hypothetical line along which items are located. Items falling close to this line contribute to the single dimension being examined, while those that fall far from the line are discarded since these items indicate construct-irrelevant variance. The fit of the final 25-item L-QoL to the Rasch model was good (overall item fit was −0.124 ± SD 0.82 and overall person fit was −0.701 ± SD 0.66). The items showed invariance of the scale across the trait.

Ability to detect change. Not reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Unlike many instruments that measure health-related QOL using multidimensional constructs that yield a profile of scores, the L-QoL provides a single unidimensional score and is based on the needs-based model of QOL. Although testing in the original development and validation study showed good reliability and validity, additional testing is required to confirm these initial findings. In particular, the original validation study examined construct validity in relation to a self-report measure of disease activity (flare) and a nonvalidated self-reported measure of disease severity. Administration of the instrument to a clinical cohort wherein physician-assessed measures of both disease activity and damage are available will yield further insight into both construct validity and also discriminant validity, or the independence of the L-QoL from other disease assessments in SLE. In addition, information on responsiveness was not available and will be needed to assess whether the measure might be applied to treatment studies of SLE. Finally, validation of the instrument in other populations, including patients with more severe disease phenotypes, will be useful.

AUTHOR CONTRIBUTIONS

Dr. Yazdany drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Yazdany J, Yelin E. Health-related quality of life and employment among persons with systemic lupus erythematosus. Rheum Dis Clin North Am 2010;36:15–32, vii.
2. Strand V, Gladman D, Isenberg D, Petri M, Smolen J, Tugwell P. Endpoints: consensus recommendations from OMERACT IV. Outcome Measures in Rheumatology. Lupus 2000;9:322–7.
3. US Department of Health and Human Services Food and Drug Administration. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2009. URL: http://www.fda.gov.laneproxy.stanford.edu/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.
4. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36) I: conceptual framework and item selection. Med Care 1992;30:473–83.
5. Jolly M. How does quality of life of patients with systemic lupus erythematosus compare with that of other common chronic illnesses? J Rheumatol 2005;32:1706–8.
6. Leong KP, Kong KO, Thong BY, Koh ET, Lian TY, Teh CL, et al. Development and preliminary validation of a systemic lupus erythematosus-specific quality-of-life instrument (SLEQOL). Rheumatology (Oxford) 2005;44:1267–76.
7. Kuriya B, Gladman DD, Ibanez D, Urowitz MB. Quality of life over time in patients with systemic lupus erythematosus. Arthritis Rheum 2008;59:181–5.
8. Moses N, Wiggers J, Nicholas C, Cockburn J. Prevalence and correlates of perceived unmet needs of people with systemic lupus erythematosus. Patient Educ Couns 2005;57:30–8.
9. McElhone K, Abbott J, Shelmerdine J, Bruce IN, Ahmad Y, Gordon C, et al. Development and validation of a disease-specific health-related

quality of life measure, the LupusQol, for adults with systemic lupus erythematosus. Arthritis Rheum 2007;57:972–9.

10. McElhone K, Castelino M, Abbott J, Bruce IN, Ahmad Y, Shelmerdine J, et al. The LupusQoL and associations with demographics and clinical measurements in patients with systemic lupus erythematosus. J Rheumatol 2010;37:2273–9.
11. Jolly M, Pickard SA, Mikolaitis RA, Rodby RA, Sequeira W, Block JA. LupusQoL-US benchmarks for US patients with systemic lupus erythematosus. J Rheumatol 2010;37:1828–33.
12. Gonzalez-Rodriguez V, Peralta-Ramirez MI, Navarrete-Navarrete N, Callejas-Rubio JL, Santos Ruiz AM, Khamashta M. Adaptation and validation of the Spanish version of a disease-specific quality of life measure in patients with systemic lupus erythematosus: the Lupus quality of life. Med Clin (Barc) 2010;134:13–6. In Spanish.
13. Jolly M, Pickard AS, Wilke C, Mikolaitis RA, Teh LS, McElhone K, et al. Lupus-specific health outcome measure for US patients: the LupusQoL-US version. Ann Rheum Dis 2010;69:29–33.
14. Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas 1960;20:141–51.
15. Leong KP, Chong EY, Kong KO, Chan SP, Thong BY, Lian TY, et al. Discordant assessment of lupus activity between patients and their physicians: the Singapore experience. Lupus 2010;19:100–6.
16. Freire EA, Bruscato A, Leite DR, Sousa TT, Ciconelli RM. Translation into Brazilian Portuguese, cultural adaptation and validation of the systemic lupus erythematosus quality of life questionnaire (SLEQOL). Acta Reumatol Port 2010;35:334–9.
17. Kong KO, Ho HJ, Howe HS, Thong BY, Lian TY, Chng HH, et al, for the Tan Tock Seng Hospital Systemic Lupus Erythematosus Study Group. Cross-cultural adaptation of the Systemic Lupus Erythematosus Quality of Life Questionnaire into Chinese. Arthritis Rheum 2007;57:980–5.
18. Leong KP, Kong KO, Howe HS. LupusQoL, a new systemic lupus erythematosus–specific quality of life measure: comment on the article by McElhone et al [letter]. Arthritis Rheum 2008;59:1047–8.
19. McElhone K, Teh LS, Abbott J. Reply [letter]. Arthritis Rheum 2008; 59:1048–9.
20. Doward LC, McKenna SP, Whalley D, Tennant A, Griffiths B, Emery P, et al. The development of the L-QoL: a quality-of-life instrument specific to systemic lupus erythematosus. Ann Rheum Dis 2009;68: 196–200.

Summary Table for Health-Related Quality of Life (HRQOL) Measures in Adult Systemic Lupus Erythematosus (SLE)*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
LupusQoL	HRQOL measure in adult SLE; 34 items, 8 domains (physical health, emotional health, body image, pain, planning, fatigue, intimate relationships, burden to others)	Patient-completed written or electronic questionnaire	<10 minutes	<5 minutes to score	Mean raw score transformed to scores ranging from 0 = worst HRQOL to 100 = best HRQOL	Good internal consistency (Cronbach's α = 0.88–0.96), good test–retest reliability (ICC 0.72–0.93).	Content validity based on patients generating items and providing feedback, reasonable concurrent validity (with SF-36) and discriminant validity (functions independently from disease activity or damage); limited construct validity testing (more disease activity generally associated with poorer HRQOL)	Not reported	Translations available in numerous languages, rigorous development and initial validation methods, additional psychometric testing has also been performed in US and Spanish populations	Studies evaluating responsiveness are needed, factor structure requires further investigation
SLEQOL	HRQOL measure in adult SLE; 40 items, 6 domains (physical functioning, activities, symptoms, treatment, mood, self-image)	Patient-completed written questionnaire	<5 minutes	Not reported	Score range 40–280; higher values mean worse QOL	Good internal consistency (Cronbach's α = 0.95 for summary score, but varied from 0.76–0.93 for domains), test–retest reliability was variable (ICC 0.83 for summary score but 4 domains had ICC <0.6)	Content validity assessed by eliciting patient feedback for items originally developed by health professionals, low concurrent validity (with SF-36), good discriminant validity (with SLAM, SLEDAI, SDI), construct validity analysis limited (score varied with self-perceived changes in global QOL)	Multiple techniques (including SRM and RE) demonstrated better responsiveness than SF-36. MCID was ~25	Only measure with published information regarding responsiveness and MCID	Reliability of individual domains is only moderate, concurrent validity with SF-36 is poor, floor effects demonstrated
L-QoL	Unidimensional needs-based assessment of QOL in SLE	Patient-completed written questionnaire	<5 minutes	Not reported	Score range 0–25; higher scores indicate worse QOL	Good internal consistency (Person-separation reliability 0.91–0.92), test–retest reliability good (ICC 0.95)	Content validity based on items being derived from patient interviews, Rasch analysis employed, construct validity supported by associations with self-reported disease activity and damage in SLE as well as employment outcomes, concurrent validity with Nottingham Health Profile scores	Not reported	Provides a single unidimensional score and initial validation study demonstrates good psychometric properties	Additional validation needed, including administration to clinical cohorts with more severe disease to allow assessment of the measure's relationship with physician assessed disease activity and damage, and evaluation of responsiveness

* LupusQoL = Lupus Quality of Life; ICC = intraclass correlations; SF-36 = Medical Outcomes Study Short Form 36; SLEQOL = Systemic Lupus Erythematosus Quality of Life; SLAM = Systemic Lupus Erythematosus Activity Measure; SLEDAI= Systemic Lupus Erythematosus Disease Activity Index; SDI = Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index; SRM = standardized response mean; RE = relative efficacy; MCID = minimum clinically important difference; L-QoL = Lupus Quality of Life.

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S47–S58
DOI 10.1002/acr.20575
© 2011, American College of Rheumatology

Measures of Symptoms and Disease Status in Ankylosing Spondylitis

Ankylosing Spondylitis Disease Activity Score (ASDAS), Ankylosing Spondylitis Quality of Life Scale (ASQoL), Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), Bath Ankylosing Spondylitis Functional Index (BASFI), Bath Ankylosing Spondylitis Global Score (BAS-G), Bath Ankylosing Spondylitis Metrology Index (BASMI), Dougados Functional Index (DFI), and Health Assessment Questionnaire for the Spondylarthropathies (HAQ-S)

JANE ZOCHLING

INTRODUCTION

Outcome measurement in spondylarthritis, particularly ankylosing spondylitis (AS) has been a rapidly growing field over the last decade, with enormous progress being made in patient-reported outcomes, clinical assessments, physical measurements and composite scoring of disease state, and response to treatment. Many of these advances arose out of need, when anti–tumor necrosis factor therapies were found to have a role in the treatment of AS patients and therefore required appropriate clinical assessment. The Assessment of SpondyloArthritis international Society (ASAS) was first formed in 1995 as a group of clinicians and methodologists with a shared interest in outcome measurement in AS patients, and has grown to incorporate early diagnosis, classification, development and validation of outcome measures, and evaluation of therapeutic modalities.

The instruments reviewed here include those recommended in the ASAS core sets for clinical record keeping (in daily clinical practice) and for clinical research, as over time these have been extensively validated and implemented across different clinical settings. The core sets describe those health-related domains that "should" be measured in AS patients in different settings, and recommend appropriate instruments that can be used for that domain. Additional measures included are the AS Quality of Life scale and the Health Assessment Questionnaire for the Spondylarthropathies, which cover health domains not included in the original core sets but have been shown to be important to AS patients through the World Health Organization International Classification of Functioning, Disability and Health projects (1). Finally, the AS Disease Activity Score has also been presented, as an alternative to the Bath Ankylosing Spondylitis Disease Activity Index, as one of the newest measures constructed to assess disease activity.

Jane Zochling, MBBS, MMed, PhD: Menzies Research Institute Tasmania, Australia.

Address correspondence to Jane Zochling, MBBS, MMed, PhD, Private Bag 23, Hobart, Tasmania, Australia. E-mail: Jane.Zochling@utas.edu.au.

Submitted for publication June 12, 2011; accepted in revised form July 28, 2011.

ANKYLOSING SPONDYLITIS DISEASE ACTIVITY SCORE (ASDAS)

Description

Purpose. To measure disease activity in ankylosing spondylitis (AS) based on a composite score of domains relevant to patients and clinicians, including both self-reported items and objective measures. The ASDAS was first published in 4 draft forms in 2008 (2), and 2 final working versions for use in patients with AS were selected by the Assessment of SpondyloArthritis international Society (ASAS) membership (2,3).

Content. The score includes patient-reported assessments of back pain, duration of morning stiffness, peripheral joint pain and/or swelling, general well-being, and a serologic marker of inflammation (erythrocyte sedimentation rate [ESR] or C-reactive protein [CRP]). The ASDAS including CRP has been presented as the preferred version and the one including ESR as the alternative version.

Number of items. Five items are combined to give a single disease activity score.

Response options/scale. Continuous scale from zero with no defined upper end.

Endorsements. The ASDAS has been endorsed by the ASAS and by Outcome Measures in Rheumatology (4).

Examples of use. The ASDAS has been validated in several observational cohorts and trial populations (5) and is now being used regularly in clinical trials, including longitudinal studies of spondylarthritis patients receiving tumor necrosis factor (TNF) inhibitors (6,7).

Practical Application

How to obtain. The ASDAS and aids for its calculation are available online at http://www.asas-group.org.

Method of administration. Patient response items and a serologic measure of inflammation are mathematically combined to give the ASDAS.

Scoring. The score is most easily calculated using an online calculator or a hand-held calculator, although it is possible to calculate longhand or using a quick ASDAS calculation form available online.

Score interpretation. Score ranges from zero (reflecting no disease activity) with the upper end of the scale being determined by the level of the CRP or ESR.

Respondent burden. The time for the patient to complete the items is very short (estimated at <1 minute) with only 4 single-item questions.

Administrative burden. Scoring is fast with access to the online calculator, hand-held calculator, or quick assessment forms, but is unwieldy without such a tool.

Translations/adaptations. The ASDAS has been used in axial spondylarthritis (7) and psoriatic arthritis (8) to date. Since the questions used are taken from the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), there are similar translations available.

Psychometric Information

Method of development. Items were generated by the 60 ASAS members using a 3-round Delphi process. There were no patients directly involved in item generation. The refined score was then derived using a 3-step statistical approach (principal component analysis, discriminant function analysis, and linear regression analysis) using data from the International Study on Starting TNF Blocking Agents in AS (9). Cross-validation was carried out in the independent Outcome in Ankylosing Spondylitis International Study database.

Acceptability. The ASDAS is easy to understand, but requires availability of both patient-reported outcomes and serologic values.

Reliability. Not reported.

Validity. Regarding content validity, the score items were generated by an international expert group of rheumatologists (ASAS) interested in AS, and the inclusion of serologic markers of inflammation improves the face validity over solely patient-reported domains. Extensive statistical analysis has minimized redundancy between items. Regarding construct validity, Pearson's correlations between the ASDAS-CRP (ASDAS-ESR) and patient global assessment was 0.74 (0.71), and with physician global assessment 0.47 (0.54). The ASDAS shows excellent discrimination between high and low disease activity states as defined by the physician global assessment (standardized mean difference [SMD] at baseline 1.33 for ASDAS-CRP and SMD 1.55 for ASDAS-ESR) based on the Norwegian disease-modifying antirheumatic drug (NOR-DMARD) database (10), and between patients treated with TNF blockers and with placebo (SMD 1.50 for ASDAS-CRP and SMD 1.51 for ASDAS-ESR) based on participants in randomized controlled trials of TNF blockers for AS (3).

Ability to detect change. The ASDAS was sensitive to improvement with TNF inhibitors in patients with axial spondylarthritis, with an effect size (ES) of 2.04 and a standardized response mean (SRM) of 1.45, and was more responsive than BASDAI (ES 1.86, SRM 1.36) (7). The ASAS group defined 4 important disease states by consensus: inactive disease, moderate, high, and very high disease activity, and relevant cut offs between these states were calculated from the NOR-DMARD database at 1.3, 2.1, and 3.5 units, respectively. Clinically important improvement was found to be 1.1 units or greater and major improvement was defined as a change of 2.0 units or more (11).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The ASDAS is still being validated, but is emerging as the best measure of disease activity in AS on the basis of including both patient-generated items and objective measures of inflammation and on having, to date, equivalent or superior performance when compared to the BASDAI.

Caveats and cautions. Further validation is required both in AS and in other patient groups to better understand the ASDAS and how it performs as a measure of disease activity, particularly with regard to reproducibility. Although endorsed by ASAS, it is not yet included in any of the ASAS measurement core sets (12,13).

Clinical usability. The ASDAS is simple enough to use in the clinical setting, but more evaluation is required as to its psychometric properties.

Research usability. Initial psychometric evaluation supports the use of the ASDAS in a research setting, without appreciable burden on either respondent or administrator.

ANKYLOSING SPONDYLITIS QUALITY OF LIFE SCALE (ASQoL)

Description

Purpose. To measure the impact of ankylosing spondylitis (AS) on health-related quality of life from the patient's perspective. The original instrument was published in 2003 for use in AS patients, and has not been updated (14).

Content. The questionnaire includes items related to the impact of disease on sleep, mood, motivation, coping, activities of daily living, independence, relationships, and social life.

Number of items. 18.

Response options/scale. Yes/no responses.

Endorsements. None.

Examples of use. The ASQoL is the most frequently used disease-specific measure of health-related quality of life in AS studies. It has recently been used in population studies (15) and to assess the effect of anti–tumor necrosis factor therapy in AS patients (16).

Practical Application

How to obtain. Available online at http://www.asas-group.org.

Method of administration. Self-report.

Scoring. Dichotomous responses, with 0 scored for a "no" and 1 scored for a "yes" for each item. Total score is the sum of the individual responses.

Score interpretation. Score range is 0–18, with higher scores reflecting greater impairment of health-related quality of life.

Respondent burden. Between 2 and 16 minutes (median 4 minutes) to complete.

Administrative burden. Less than a minute to score.

Translations/adaptations. The original UK English ASQoL continues to be translated and validated in other languages, including US English, Canadian French and English, German, Italian, Spanish, Swedish (17), French (17,18), Chinese (19), Hungarian (20) and Turkish (17,21). It has also been validated in patients with axial spondylarthritis (22).

Psychometric Information

Method of development. Items were generated from patient interviews in the UK and the Netherlands, and scaling properties were tested using Rasch analyses. Conceptually, the ASQoL is based on a needs-based model of health.

Acceptability. The ASQoL is readable and simple to complete.

Reliability. Regarding internal consistency, Cronbach's α is reported between 0.89–0.92 in the different study groups (UK, Netherlands, time 1 or time 2). Test–retest reliability was r = 0.91–0.92, and intraclass correlation coefficients were 0.91 (Netherlands) and 0.92 (UK).

Validity. For content validity, the measure was derived from patient interviews, imparting high relevance to AS patients. Regarding construct validity, the ASQoL correlates moderately well with other AS-specific health outcome measures, including the Nottingham Health Profile components (Spearman's rank correlation coefficient for physical mobility 0.78–0.79, energy 0.73–0.74, pain 0.79–0.81, emotional reactions 0.72–0.73, sleep 0.54 –0.59, and social isolation 0.50–0.53), Bath Ankylosing Spondylitis Functional Index (correlation coefficient 0.72–0.75), Leeds Disability Questionnaire (0.70), Dougados Functional Index (0.80) (14), and the Bath Ankylosing Spondylitis Disease Activity Index (0.79) (23).

Ability to detect change. The modified standardized response mean measured against AS health transition is reported as −0.35 for improvement and 0.57 for deterioration in health ($P < 0.01$) and, measured against general health transition, is −0.73 for improvement and 0.48 for deterioration (23). Minimum clinically important difference (MCID) has not been reported for ASQoL, but patient acceptable symptom state has been calculated at 8.0 (24).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The ASQoL is a validated disease-specific health-related quality of life measure for patients with AS, which captures important information on limitation of activities and participation that are not covered in other outcome measures.

Caveats and cautions. Information on cut offs and MCID is currently lacking.

Clinical usability. Psychometric evaluation supports the use of the ASQoL in a clinical setting.

Research usability. Psychometric evaluation supports research use, with minimal administrative or respondent burden.

BATH ANKYLOSING SPONDYLITIS DISEASE ACTIVITY INDEX (BASDAI)

Description

Purpose. To measure patient-reported disease activity in patients with ankylosing spondylitis (AS). The instrument was first published in 1994 (25) using visual analog scales.

Content. The index includes patient-reported levels of back pain, fatigue, peripheral joint pain and swelling, localized tenderness, and the duration and severity of morning stiffness.

Number of items. 6 items.

Response options/scale. Numeric response scale (0–10) or visual analog scale (VAS, 0–10 cm) anchored by adjectival descriptors "none" and "very severe." Duration of morning stiffness is anchored by a time scale (0–2 or more hours).

Endorsements. The BASDAI has been endorsed by the Assessment of SpondyloArthritis international Society (ASAS) for the measurement of disease activity (13).

Examples of use. The BASDAI has been the most frequently used measure of disease activity in clinical trials (26–28) and is recommended to assess response to anti–tumor necrosis factor therapies in AS patients (29,30). Many regulatory bodies require the BASDAI to be reported for prescribing purposes.

Practical Application

How to obtain. Available online (in multiple translations) at http://www.asas-group.org.

Method of administration. Patient self-report questionnaire.

Scoring. The scores for questions 5 and 6 (severity and duration of morning stiffness) are averaged, the result is then averaged with the remaining 4 question scores to give a final score out of 10.

Score interpretation. Ranges from 0 (no disease activity) to 10 (maximal disease activity). A cut off of 4 is used to define active disease (29–31). Reference percentile charts have been published (32).

Respondent burden. Time to complete is between 30 seconds and 2 minutes (mean 67 seconds).

Administrative burden. Scoring requires less than a minute.

Translations/adaptations. There are validated translations available in more than 20 languages, including English (25), French (33), Swedish (34), Dutch (35), Turkish (36), German (37,38), Arabic (39), Spanish (40), and Portuguese (41).

Psychometric Information

Method of development. Items were generated based on the expert opinion of a group of physiotherapists, researchers, rheumatologists, and patient input.

Acceptability. The BASDAI is understandable, but missing data have been reported in up to 20% of completions (23). There have been no floor or ceiling effects reported.

Reliability. Internal consistency is good with a Cronbach's α of 0.84–0.87 (23,42). Test–retest reliability was good when assessed for inpatients over a 24-hour period (r = 0.93, $P < 0.001$) (25), and when assessed by postal survey in 162 AS patients who reported no change on an AS-specific health transition question over a 1-week period (intraclass correlation coefficient of 0.87 (95% confidence interval [95% CI] 0.83–0.91) (23).

Validity. Regarding content validity, the measure was developed by experts in the field with patient input, reflecting items relevant to both patients and clinicians. For construct validity, the BASDAI correlated well with the earlier Bath Disease Activity Index with no significant differences in score distribution, reproducibility, or sensitivity. There is good correlation with the Ankylosing Spondylitis Quality of Life questionnaire (Pearson's correlation coefficient 0.79) and BASDAI is significantly higher in AS patients unable to work due to ill health ($P < 0.01$) (23). Although largely validated using the VAS, it should be noted that ASAS prefers the use of numeric rating scales (NRS), and there is evidence supporting the replacement of the original VAS answer modalities with NRS (43) in patient self-report instruments in AS.

Ability to detect change. The modified standardized response mean measured against AS health transition is reported as −0.74 for improvement and 0.60 for deterioration in health ($P < 0.01$), and measured against general health transition is −1.02 for improvement and 0.74 for deterioration (23). The minimum clinically important difference from the patient's perspective has been reported as 10 mm or 22.5% with a sensitivity of 0.65 and specificity of 0.82, determined using receiver operating characteristic curves analyses (44). A 50% improvement in BASDAI with an intervention (BASDAI50) has been defined as a response to that intervention. Initial evaluation of the patient acceptable symptom state (PASS) in AS patients (45) found the mean change in BASDAI over 12 weeks for the patient to feel well was −3.5 (SD 2.3), and there was significant correlation between the BASDAI 50% responders and patients achieving PASS. Tubach et al described a PASS cut off for BASDAI of 34.5 mm (95% CI 30.9–38.9 mm) (46).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BASDAI has been the preferred instrument for measuring disease activity in AS patients since its development, and has become the gold standard measure in clinical trials and in daily patient care, specifically in regard to prescribing anti–tumor necrosis factor therapies. It is responsive, easy to administer, and easy to interpret.

Caveats and cautions. As a patient-generated index, the BASDAI does not include any objective measures that might relate to disease activity, and it does not include the clinician's perspective. Scores are dependent on what the patient perceives as being related to their AS.

Clinical usability. The BASDAI is easy to use and has found its way into daily clinical practice, although internal consistency (Cronbach's $\alpha = 0.87$) is a little lower than the 0.90 considered important for an instrument's use in individual patients.

Research usability. The ease of use, the reproducibility, and the sensitivity to change seen with the BASDAI make it a useful instrument for research purposes.

BATH ANKYLOSING SPONDYLITIS FUNCTIONAL INDEX (BASFI)

Description

Purpose. To define and monitor physical functioning in patients with ankylosing spondylitis (AS). The index was developed in 1994 (47) using visual analog scales.

Content. Eight items concerning activities referring to the functional anatomy of the patients (bending, reaching, changing position, standing, turning, and climbing steps), and 2 items assessing the patients' ability to cope with everyday life.

Number of items. 10 items.

Response options/scale. Numeric response scale (0–10) or visual analog scale (0–10 cm) anchored by adjectival descriptors "easy" and "impossible."

Endorsements. Endorsed by the Assessment of SpondyloArthritis international Society.

Examples of use. The BASFI is the most widely used functional index for assessment of AS patients, primarily in studies of disease impact (48,49) and in clinical trials (26,27).

Practical Application

How to obtain. Available online at http://www.asas-group.org.

Method of administration. Patient self-report questionnaire.

Scoring. The mean of the individual scores is calculated to give the overall index score.

Score interpretation. Score range is 0–10, with 0 reflecting no functional impairments and 10 reflecting maximal impairment. Reference percentile charts have been published (32).

Respondent burden. The instrument takes <3 minutes to complete.

Administrative burden. Scoring is simple and quick.

Translations/adaptations. The BASFI has been translated and validated in over 18 languages, including English (47), French, German, Dutch, Spanish (40), Portuguese (41), and Chinese (50).

Psychometric Information

Method of development. Items were generated by an expert group of physiotherapists, researchers, rheumatologists, and patients. No factor analysis or principal component analysis was performed.

Acceptability. The questions are easy to comprehend. The median score has been reported as 2.0, with clustering at the lower end of the scale (51); studies with item-response theory have indicated that the BASFI is not a linear scale (51,52).

Reliability. Internal consistency is excellent, with Cronbach's α reported at 0.936 (53). Test–retest reliability was good when scores were repeated after a 24-hour interval ($r = 0.89$, $P < 0.001$) (47), and in patients stable on anti–tumor necrosis factor treatment ($r = 0.92$, $P < 0.0001$) (54). Interobserver reliability measured between patient self-report and the score given by a physiotherapist after observing the tasks being performed has also been reported ($r = 0.87–0.89$, $P < 0.001$).

Validity. Regarding content validity, the BASFI was developed by a multidisciplinary group of experts in AS, with input from AS patients. For construct validity, there is good evidence of validity through comparison with instruments that measure similar or related constructs, and with measures of mobility (55). The BASFI is highly correlated with the Dougados Functional Index (Spearman's correlation coefficient 0.89) (35). Correlations are less strong with patient-reported disease activity ($r = 0.33$) or physician-reported disease activity ($r = 0.33$) (35).

Ability to detect change. Responsiveness statistics for the BASFI have been published as an effect size (ES) of 0.36 (moderate) and standardized response mean (SRM) of 0.46 for improvement, and an ES of 0.70 and SRM of 0.72 for deterioration in the setting of a trial of nonsteroidal antiinflammatory drugs. Although responsive in placebo-controlled trials of active drugs (56,57), the BASFI is less responsive in the setting of physical therapy interventions (58,59). The minimum clinically important difference from the patient's perspective has been reported as 7 mm or 17.5% with a sensitivity of 0.60 and specificity of 0.85, determined using receiver operating characteristic curves analyses (44). In assessing patient acceptable symptom state (PASS), Dougados et al found the mean change in BASFI over 12 weeks for the patient to feel well was somewhat larger at −2.4 (SD 2.0) (45), with a PASS cut off reported by Tubach et al of 31.4 mm (95% confidence interval 26.9–37.0 mm) (46), and slightly higher by Maksymowych et al using the numeric rating scale of 4.0 cm (24).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BASFI is a valid measure of physical function in AS patients and has good discrimination between groups and interventions. It is simple and easy to use and score. The tool was designed for use in AS, but can also be used in the other spondylarthritides.

Caveats and cautions. The BASFI may not be sufficiently sensitive to detect subtle changes in functioning in the relatively well AS patient, or in trials of physical therapies.

Clinical usability. Psychometric evaluation supports its use in a clinical setting.

Research usability. Psychometric evaluation supports its use in a research setting, being short, easy to complete, reproducible, and sensitive to change at a group level.

BATH ANKYLOSING SPONDYLITIS GLOBAL SCORE (BAS-G)

Description

Purpose. To give a global assessment of the well-being of the person with ankylosing spondylitis (AS) over a given time period. It was first introduced by Jones et al in 1996 (60).

Content. Two visual analog scales to measure the effect of AS on the respondent's well-being, the first estimated over the last week, the second over the last 6 months.

Number of items. Two items.

Response options/scale. Scale between 0 (none) to 10 (very severe effect).

Endorsements. Endorsed by Assessment of SpondyloArthritis international Society (ASAS).

Examples of use. Relevant references in which instrument has been used.

Practical Application

How to obtain. Available online at http://www.asas-group.org.

Method of administration. Patient-completed.

Scoring. There are no specific scoring instructions.

Score interpretation. Higher scores reflect a more severe effect of the disease on general well-being.

Respondent burden. Less than a minute required to complete.

Administrative burden. Minimal.

Translations/adaptations. Few published translations have been validated; Dutch and Norwegian translations are available on the ASAS web site.

Psychometric Information

Method of development. Not reported.

Acceptability. Very simple to understand and complete.

Reliability. There is no evidence for internal consistency. Test–retest reliability was good when scores were repeated after a 24-hour interval ($r = 0.84$ for 1 week, $r = 0.93$ for 6 months) in the original study of 329 AS patients (60), and in patients stable on anti–tumor necrosis factor treatment over 1 week ($r = 0.74$, $P < 0.0001$) (54).

Validity. Regarding construct validity, the BAS-G correlated better with other patient-reported health measures (r = 0.73 for Bath Ankylosing Spondylitis Disease Activity Index and r = 0.30–0.59 for Bath Ankylosing Spondylitis Functional Index) than it did with objective physical measures (r = −0.16 for Bath Ankylosing Spondylitis Metrology Index).

Ability to detect change. Satisfactory sensitivity to change was reported, with a mean difference between pre- and postglobal scores of 1.54, SEM 0.31, $P = 0.001$ (60). The minimal clinically important difference from the patient's perspective has been reported as 15 mm or 27.5% with a sensitivity of 0.61 and specificity of 0.74, determined using receiver operating characteristic curve analyses (44).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BAS-G is the simplest of the patient-reported outcome measures in AS and a good indicator of the patient's perspective on the overall effect of their health on well-being. It is appropriate for assessing interventions aiming to improve overall disease impact, and likely reflects many constructs that are implicitly relevant for patients including fatigue, emotions, fears and anxiety, side effects of medications, and restrictions in social roles. The BAS-G is easily applied to all spondylarthritis groups.

Caveats and cautions. Being unidimensional, the BAS-G is entirely reliant on the patient's perception of their disease and how it affects the aspects of their life that are important to them. Patients with similar physical functioning may express the impact of disease differently. The BAS-G has not been as well evaluated as the other Bath AS indices.

Clinical usability. The BAS-G is simple to use, simple to evaluate, and appropriate for use in the clinical setting.

Research usability. There is insufficient psychometric evidence to support widespread research use.

BATH ANKYLOSING SPONDYLITIS METROLOGY INDEX (BASMI)

Description

Purpose. To quantify the mobility of the axial skeleton in ankylosing spondylitis (AS) patients and allow objective assessment of clinically significant changes in spinal movement. The BASMI was first published in 1994 (61) as a 2-point scale, was adapted a year later into a 10-point scale, and has more recently been proposed and validated as a linear construct (62).

Content. Clinical measures of cervical rotation, tragus to wall distance, lumbar flexion, lumbar side flexion, and intermalleolar distance.

Number of items. 5 items.

Response options/scale. Each item is scored from 0–10 based on individually defined cut points. Ranges are given as cervical rotation (>85.0° to ≤8.5°), tragus to wall (<10 cm to ≥38 cm), lumbar flexion (>7.0 cm to ≤0.7 cm), lumbar side flexion (>20.0 cm to <1.2 cm), and intermalleolar distance (≥120 cm to <30 cm).

Endorsements. Endorsed by the Assessment of SpondyloArthritis international Society (ASAS).

Examples of use. The BASMI is included in the ASAS core sets as the preferred measure of spinal mobility. It has been used in clinical trials of anti–tumor necrosis factor agents in AS patients (63,26), and more recently was the outcome measure used to show that spinal mobility is determined by both spinal inflammation and by structural damage (64).

Practical Application

How to obtain. The BASMI_{10} is available at http://www.asif.rheumanet.org/basmi-10-e.pdf, and the linear version is available at http://www.asif.rheumanet.org/basmi-lin-e.pdf. Translations into Norwegian, German, and Danish are also available at http://www.asas-group.org and http://www.asif.rheumanet.org/assessment.htm.

Method of administration. Measurements are performed by health care providers who have been trained to perform the clinical examinations required.

Scoring. In the original instrument (BASMI_2) (61), each continuous assessment was converted into a nominal score of 0, 1, or 2. The next year a second nominal version was published (BASMI_{10}) (65), with individual assessments scored between 0 and 10. More recently a linear version has been proposed ($\text{BASMI}_{\text{lin}}$) (62), with scoring ranges similar to the BASMI10.

Score interpretation. Individual scores are summed for the BASMI_2 or averaged for the BASMI_{10} to give a final score between 0 and 10, where a higher score reflects more significant impairment of spinal mobility. Normative values have been published previously (32) using the BASMI_{10}.

Respondent burden. The BASMI takes ~5–10 minutes to complete, depending on the experience of the clinician.

Administrative burden. Scoring is straightforward and takes less than a minute to complete.

Translations/adaptations. Multiple languages are available, including English (61,65), Finnish (66), Portuguese (67), German, Danish, and Norwegian.

Psychometric Information

Method of development. Measurements were chosen by a group of rheumatologists, physiotherapists, and research associates with a special interest in AS, based on an extensive literature review and clinical experience. The chosen measurements were found to be most reliable and clinically useful to reflect axial "status."

Acceptability. The instrument is easily understood and is published with step-by-step instructions as to how one should perform the clinical measures. Most of the measures are easy to carry out, although appropriate measurement of intermalleolar distance can be difficult in a small clinic room. The BASMI_2 does not perform as well as the BASMI_{10} or the $\text{BASMI}_{\text{lin}}$, scoring lower at the lower end of the scale and higher at the higher end, with different magnitude of changes at different ends of the scale. The BASMI_{10} and $\text{BASMI}_{\text{lin}}$ both behave in a more linear fashion, making them more suitable for monitoring AS patients over time (62).

Reliability. Interrater reliability has been shown to be good; comparisons between 3 physiotherapists revealed the following: cervical rotation ($r = 0.98$, $P < 0.001$), tragus to wall ($r = 0.99$, $P < 0.001$), lumbar side flexion ($r = 0.94$, $P < 0.001$), lumbar flexion as measured by the modified Schober's method ($r = 0.99$, $P < 0.001$), and intermalleolar distance ($r = 0.98$, $P < 0.001$). Intraobserver reliability for the same 3 physiotherapists on consecutive days showed similar high values: cervical rotation ($r = 0.99$, $P < 0.001$), tragus to wall ($r = 0.99$, $P < 0.001$), lumbar side flexion ($r = 0.98$, $P < 0.001$), lumbar flexion as measured by the modified Schober's method ($r = 0.99$, $P < 0.001$), and intermalleolar distance ($r = 0.99$, $P < 0.001$).

Validity. Content validity is fair, with the initial instrument development based on an extensive literature review and a panel of clinicians and research associates with a special interest in AS. Thoracic spinal mobility is underrepresented in this instrument as indicated by the lack of association between the BASMI and thoraco-abdominal motion in AS patients (68), and the ASAS group recommends the addition of chest expansion to the core set for clinical evaluation in AS patients to address this limitation. Regarding construct validity, the BASMI has been shown to discriminate between patients with and without radiographic change due to AS (69). The BASMI does not correlate strongly with changes in functional outcomes as measured by the Bath Ankylosing Spondylitis Functional Index ($r = 0.44$ for BASMI_2, $r = 0.46$ for BASMI_{10}, $P < 0.001$) (70). Spinal mobility, as measured by the $\text{BASMI}_{\text{lin,}}$ correlates with radiographic change as measured by the Stoke Ankylosing Spondylitis Spine Score (Spearman's $\rho = 0.6$) and less strongly with inflammation as measured on magnetic resonance imaging ($\rho = 0.3$), with both contributing independently to spinal mobility (64). For criterion validity, the comparison between the 5 BASMI measures and total scores of 20 clinical measurements (total metrology score) was good ($r = 0.92$, $P < 0.001$).

Ability to detect change. Jenkinson et al (61) reported a 30% improvement in BASMI scores over a 3-week period of inpatient treatment in 56 patients. Clinimetric properties of the 3 versions of BASMI were tested in 187 patients from the Outcomes in Ankylosing Spondylitis International Study, giving calculated Guyatt's effect sizes of 0.66 for BASMI_2, 0.95 for BASMI_{10}, and 1.04 for the $\text{BASMI}_{\text{lin}}$ (62).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The BASMI is valid, reproducible, and easy to perform with minimal training, giving valuable information about spinal mobility due to inflammation or structural damage. It is the measure of choice put forward by ASAS to measure spinal mobility in clinical trials. It is an appropriate measure for assessing the effect of interventions in patients with AS and spondylarthritis. On the strength of our current understanding of these measures, both the BASMI_{10} and the $\text{BASMI}_{\text{lin}}$ perform well for the assessment of spinal mobility.

Caveats and cautions. The instrument does not assess thoracic mobility in isolation, and it is recommended that the BASMI be used in conjunction with chest expansion or another measure of thoracic mobility for more accurate assessment of spinal mobility. Many of the earlier published studies that refer to the BASMI do not specify if it is the BASMI_2 or the BASMI_{10} that is being carried out, which may have implications on the interpretation of results.

Clinical usability. The BASMI_{10} and $\text{BASMI}_{\text{lin}}$ are sufficiently sensitive and reliable for use in clinical practice; however, the administrative burden of up to 10 minutes to carry out the physical measures may limit its use in this setting.

Research usability. The BASMI lends itself well to use in clinical trials, is sensitive enough to detect change between treatment groups, and adds important information not obtained with other outcome measures.

DOUGADOS FUNCTIONAL INDEX (DFI)

Description

Purpose. To assess the functional abilities of persons with ankylosing spondylitis (AS). The DFI was first published as the Ankylosing Spondylitis Functional Index in 1988 (71). A more recent modification uses a 5-point Likert response scale in place of the original 3-point scale described below.

Content. Items cover activities of daily living including dressing, bathing, standing, climbing stairs, changing position, bending, doing housework or usual job, coughing or sneezing, and breathing deeply. Each question begins with "Can you . . . ."

Number of items. 20.

Response options/scale. The Likert scale is 0 = yes, with no difficulty; 1 = yes, but with difficulty; and 2 = impossible to do. Item scores are added to give a total functional index.

Endorsements. Endorsed by the Assessment of SpondyloArthritis international Society as an alternative to the Bath Ankylosing Spondylitis Functional Index (BASFI) for measuring the core concept of physical function in AS.

Examples of use. The DFI continues to be used to measure physical function in studies of disease outcome (72) and to measure change in physical function due to treatment with antiinflammatory drugs in randomized controlled trials (57,73).

Practical Application

How to obtain. French, Finnish, and Russian versions are available online at www.asas-group.org.

Method of administration. Self-report questionnaire.

Scoring. The individual scores for each item are summed for a final index score.

Score interpretation. Score ranges from 0–40, with higher values reflecting higher functional impairment (worse physical functioning).

Respondent burden. Time to complete not given. Considerably longer than the BASFI.

Administrative burden. Hand-scored, time to score minimal.

Translations/adaptations. The original French version has been translated into multiple languages including English, German (74), Italian (75), Spanish (40) and Turkish (76). The DFI has been used (but not validated) in other spondylarthritis subgroups including reactive arthritis and psoriatic arthritis patients.

Psychometric Information

Method of development. Items were generated in an expert group of 3 rheumatologists experienced in the management of patients with AS. Patients were not involved in the development process. Preliminary component analysis was used to refine items.

Acceptability. The questions are easily understandable. There is clustering at the low (normal) end of the scale, and the response "yes, but with difficulty" covers a very wide range of functional restriction.

Reliability. Cronbach's alpha was not given. Interrater reliability has been reported as high between 2 independent observers (rheumatologists, not patient-scored) with an intraclass correlation coefficient (ICC) of 0.99. Intrarater reliability of an independent observer scoring the index on 2 occasions 1 week apart gave an ICC of 0.86. Test–retest reliability was good when scores were repeated after a 24-hour interval ($r = 0.96$, $P < 0.001$) in inpatients with AS (47).

Validity. Regarding content validity, the initial item pool was generated by 3 rheumatologists and refined using principal component analysis. Regarding construct validity, the DFI discriminates between AS inpatients and outpatients (47), AS patients with high and low disease activity (77), and AS smokers and nonsmokers (78). The DFI is highly correlated with the BASFI (Spearman's correlation coefficient 0.89) (35), and moderately correlated with the Health Assessment Questionnaire (HAQ) disability index ($r = 0.66$), the HAQ for the Spondylarthropathies ($r = 64$), and the Arthritis Impact Measurement Scales 2 ($r = 0.55$) (79). Correlations are less strong with patient-reported disease activity ($r = 0.32$), physician-reported disease activity ($r = 0.36$) (35), or physical limitations (including cervical rotation, $r = -0.23$; Schober's test, $r = -0.13$) (79). Criterion validity was assessed using a multiple regression model, including the DFI score, and clinically measured morning stiffness, number of nocturnal awakenings, chest expansion, Schober's test, hand-ground distance, self-physiotherapy, and pain. The correlation coefficient between the independent variables and the functional index (dependent variable) was $R^2 = 0.41$.

Ability to detect change. Responsiveness statistics for the DFI include an effect size (ES) of 0.30 (moderate) and standardized response mean (SRM) of 0.33 for improvement, and an ES of 0.47 and SRM of 0.59 for deterioration (51).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The DFI is a valid instrument for measuring physical functioning in AS and is responsive to change. It is most appropriate in patients with predominantly axial involvement.

Caveats and cautions. The concepts included in the DFI are centered on axial and large joint functioning, which may limit its use in individuals with significant peripheral joint or extraarticular involvement. There is a significant floor effect with clustering of responses toward the lower end of the scale, and the 3-response Likert scale is overly simplistic to capture subtle changes in functioning.

Clinical usability. There is insufficient evidence to support the use of the DFI for clinical practice. The time required to complete the measure may limit clinical use.

Research usability. The DFI is an appropriate measure for research use; however, respondent burden may limit its feasibility. Shorter, equally valid instruments (e.g., the BASFI) are likely to be favored in this setting.

HEALTH ASSESSMENT QUESTIONNAIRE FOR THE SPONDYLARTHROPATHIES (HAQ-S)

Description

Purpose. To assess the physical functioning of an individual with ankylosing spondylitis (AS). The HAQ-S was adapted from the original HAQ in 1990, incorporating issues of physical functioning and impairment specific to patients with AS (80).

Content. The measure includes items concerning dressing, arising, eating, walking, hygiene, reaching, gripping, and errands and chores taken from the disability index (DI) of the HAQ (81), and an additional 5 specific items concerning neck function and static posture (driving a car, using a rear-vision mirror, carrying heavy groceries, sitting for long periods, and working at a desk).

Number of items. There are 25 items (20 from the HAQ-DI and 5 unique to the HAQ-S).

Response options/scale. Responses are 0 = able to do with no difficulty; 1 = able to do with some difficulty; 2 = able to do with much difficulty; and 3 = unable to do. The final score range is 0–3.

Endorsements. No formal endorsements.

Practical Application

How to obtain. The original, more generic HAQ and its shorter form, the HAQ-DI, are available at http://aramis.stanford.edu/HAQ.html. The 5 extra questions are outlined in the original HAQ-S manuscript (80).

Method of administration. Self-report questionnaire.

Scoring. The responses are hand-scored, and the 10 individual subscale scores are averaged to give a summary HAQ-S score.

Score interpretation. The range of the final calculated summary score is 0–3, where a higher score indicates higher impairment or worse function.

Respondent burden. Time to complete is not given. Normative values are available for the original HAQ-DI (82), but not for the spondylarthritis-specific instrument.

Administrative burden. Minimal time is required for scoring.

Translations/adaptations. Multiple translations of the original English version are available, including Dutch (83,84), Finnish (85), Spanish (86), Brazilian-Portuguese (87), and Turkish (88). The HAQ-S has also been used to assess functioning in patients with psoriatic arthritis (7).

Psychometric Information

Method of development. The existing HAQ-DI was adapted for use in spondylarthritis patients by adding 5 disease-specific items, determined by the investigators (methods not given).

Acceptability. The instrument is readable and understandable. There is a floor effect with score clustering at the normal (0) end of the scale described.

Reliability. Test–retest showed stability between time 1 and time 3, with Pearson's correlation coefficient given as r = 0.92.

Validity. For content validity, the original HAQ was felt to show good face validity related to the difficulties with activities of daily living reported by a group of 300 British AS patients (80,89). Additional spondylitis-specific items were developed by the investigators. Regarding construct validity, the HAQ-S is highly correlated with the original HAQ-DI (Spearman's correlation coefficient 0.96), and moderately correlated with other measures of physical functioning, the Dougados Functional Index (r = 0.64) and the generic instrument, Arthritis Impact Measurement Scales (r = 0.80) (79). For criterion validity, correlations are less strong with physical limitations (including cervical rotation, r = −0.50; Schober's test, r = −0.36) (79).

Ability to detect change. Responsiveness statistics for the HAQ-S show an effect size (ES) of 0.20 (moderate) and standardized response mean (SRM) of 0.28 for improvement, and an ES of 0.28 and SRM of 0.72 for deterioration (51) in a cohort of AS patients treated with nonsteroidal antiinflammatory drugs.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument measures aspects of physical function and impairments in activities of daily living in patients with AS. It is sensitive in early disease as well as in advanced disease. It can be applied to all the spondylarthritides (not only AS patients).

Caveats and cautions. The HAQ-S is longer than other disease-specific measures of physical functioning and does not perform a great deal better than the alternatives. Information relating to minimum clinically important difference, relevant cut offs and patient acceptable symptom state is lacking.

Clinical usability. There is insufficient evidence to support the use of the HAQ-S in individual patient care. Respondent burden may also limit clinical use.

Research usability. There is good evidence to support the use of the HAQ-S for research use, as a valid easily administered tool to measure physical function.

AUTHOR CONTRIBUTIONS

Dr. Zochling drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Van Bechtel I, Cieza A, Boonen A, Stucki G, Zochling J, Braun J, et al. Identification of the most common problems by patients with ankylosing spondylitis using the international classification of functioning, disability and health. J Rheumatol 2006;33:2475–83.
2. Lukas C, Landewe R, Sieper J, Dougados M, Davis J, Braun J, et al. Development of an ASAS-endorsed disease activity score (ASDAS) in patients with ankylosing spondylitis. Ann Rheum Dis 2009;68:18–24.
3. Van der Heijde D, Lie E, Kvien TK, Sieper J, van den Bosch F, Listing J, et al. ASDAS, a highly discriminatory ASAS-endorsed disease activity score in patients with ankylosing spondylitis. Ann Rheum Dis 2009;68:1811–8.
4. Machado PM, Landewe RB, van der Heijde DM. Endorsement of definitions of disease activity states and improvement scores for the Ankylosing Spondylitis Disease Activity Score: Results from OMERACT 10. J Rheumatol 2011;38:1502–6.
5. Machado P, van der Heijde D. How to measure disease activity in axial spondyloarthritis? Curr Opin Rheumatol 2011;23:339–45.
6. Pedersen SJ, Sorensen IJ, Garnero P, Johansen JS, Madsen OR, Tvede N, et al. ASDAS, BASDAI and different treatment responses and their relation to biomarkers of inflammation, cartilage and bone turnover in patients with axial spondyloarthritis treated with TNFα inhibitors. Ann Rheum Dis 2011;70:1375–81.
7. Pedersen SJ, Sorensen IJ, Hermann KG, Madsen OR, Tvede N, Hansen MS, et al. Responsiveness of the Ankylosing Spondylitis Disease Activity Score (ASDAS) and clinical and MRI measures of disease activity in a 1-year follow-up study of patients with axial spondyloarthritis treated with tumour necrosis factor alpha inhibitors. Ann Rheum Dis 2010;69:1065–71.
8. Eder L, Chandran V, Shen H, Cook RJ, Gladman DD. Is ASDAS better than BASDAI as a measure of disease activity in axial psoriatic arthritis? Ann Rheum Dis 2010;69:2160–4.
9. Pham T, Landewe R, van der Linden S, Dougados M, Sieper J, Braun J, et al. An international study on starting tumour necrosis factor-blocking agents in ankylosing spondylitis. Ann Rheum Dis 2006;65: 1620–5.
10. Kvien TK, Heiberg, Lie E, Kaufmann C, Mikkelsen K, Nordvag BY, et al. A Norwegian DMARD register: prescriptions of DMARDs and biological agents to patients with inflammatory rheumatic diseases. Clin Exp Rheumatol 2005;23 Suppl 39:S188–S94.
11. Machado P, Landewe R, Lie E, Kvien TK, Braun J, Baker D, et al. Ankylosing Spondylitis Disease Activity Score (ASDAS): defining cut-off values for disease activity states and improvement scores. Ann Rheum Dis 2011;70:47–53.
12. Van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy, and clinical record keeping in ankylosing spondylitis: progress report of the ASAS Working Group. Assessments in Ankylosing Spondylitis. J Rheumatol 1999;26:951–4.
13. Zochling J, Sieper J, van der Heijde D, Braun J. Development of a core set of domains for data collection in cohorts of patients with ankylosing spondylitis receiving anti-tumor necrosis factor-alpha therapy. J Rheumatol 2008;35:1079–82.
14. Doward LC, Spoorenberg A, Cook SA, Whalley D, Helliwell PS, Kay LJ, et al. Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. Ann Rheum Dis 2003;62:20–6.
15. Bodur H, Ataman S, Rezvani A, Bugdayci DS, Cevik R, Birtane M, et al. Quality of life and related variables in patients with ankylosing spondylitis. Qual Life Res 2011;20:543–9.
16. Van der Heijde DM, Revicki DA, Gooch KL, Wong RL, Kupper H, Harnam N, et al. Physical function, disease activity, and health-related quality-of-life outcomes after 3 years of adalimumab treatment in patients with ankylosing spondylitis. Arthritis Res Ther 2009;11:R124.
17. Doward LC, McKenna SP, Meads DM, Twiss J, Revicki D, Wong RL, et al. Translation and validation of non-English versions of the Ankylosing Spondylitis Quality of Life (ASQOL) questionnaire. Health Qual Life Outcomes 2007;5:7.
18. Pham T, van der Heijde DM, Pouchot J, Guillemin F. Development and validation of the French ASQoL questionnaire. Clin Exp Rheumatol 2010;28:379–85.
19. Zhao LK, Liao ZT, Li CH, Li TW, Wu J, Lin Q, et al. Evaluation of quality of life using ASQoL questionnaire in patients with ankylosing spondylitis in a Chinese population. Rheumatol Int 2007;27:605–11.
20. Lovas K, Geher P, Whalley D, McKenna S, Meads D, Kalo Z. Hungarian

adaptation of a disease-specific quality-of-life questionnaire in patients with ankylosing spondylitis. Orv Hetil 2002;143:1893–7. In Hungarian.

21. Duruoz T, Doward LC, Cerrahoglu L, Turan Y, Yurtkuran M, Calis M, et al. Translation and validation of the Turkish version of the ankylosing spondylitis quality of life (ASQoL) Questionnaire [abstract]. Ann Rheum Dis 2008;67 Suppl II:623.
22. Jenks K, Treharne GJ, Garcia J, Stebbings S. The ankylosing spondylitis quality of life questionnaire: validation in a New Zealand cohort. Int J Rheum Dis 2010;13:361–6.
23. Haywood KL, Garratt M, Jordan K, Dziedzic K, Dawes PT. Disease-specific, patient-assessed measures of health outcome in ankylosing spondylitis: reliability, validity and responsiveness. Rheumatology (Oxford) 2002;41:1295–302.
24. Maksymowych WP, Richardson R, Mallon C, van der Heijde D, Boonen A. Evaluation and validation of the patient acceptable symptom state (PASS) in patients with ankylosing spondylitis. Arthritis Rheum 2007; 57:133–9.
25. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J Rheumatol 1994;21:2286–91.
26. Brandt J, Haibel H, Cornely D, Golder W, Gonzalez J, Reddig J, et al. Successful treatment of active ankylosing spondylitis with the anti–tumor necrosis factor α monoclonal antibody infliximab. Arthritis Rheum 2000;43:1346–52.
27. Davis JC Jr, van der Heijde D, Braun J, Dougados M, Cush J, Clegg DO, et al. Recombinant human tumor necrosis factor receptor (etanercept) for treating ankylosing spondylitis: a randomized, controlled trial. Arthritis Rheum 2003;48:3230–6.
28. Van der Heijde D, Kivitz A, Schiff MH, Sieper J, Dijkmans BA, Braun J, et al, for the Atlas Study Group. Efficacy and safety of adalimumab in patients with ankylosing spondylitis: results of a multicenter, randomized, double-blind, placebo-controlled trial. Arthritis Rheum 2006;54: 2136–46.
29. Braun J, Pham T, Sieper J, Davis J, van der Linden S, Dougados M, et al. International ASAS consensus statement for the use of anti-tumour necrosis factor agents in patients with ankylosing spondylitis. Ann Rheum Dis 2003;62:817–24.
30. Van der Heijde D, Sieper J, Maksymowych WP, Dougados M, Burgos-Vargas R, Landewe R, et al. 2010 Update of the international ASAS recommendations for the use of anti-TNF agents in patients with axial spondyloarthritis. Ann Rheum Dis 2011;70:905–8.
31. Cohen JD, Cunin P, Farrenq V, Oniankitan O, Carton L, Chevalier X, et al. Estimation of the Bath Ankylosing Spondylitis Disease Activity Index cutoff for perceived symptom relief in patients with spondyloarthropathies. J Rheumatol 2006;33:79–81.
32. Taylor AL, Balakrishnan C, Calin A. Reference centile charts for measures of disease activity, functional impairment, and metrology in ankylosing spondylitis. Arthritis Rheum 1998;41:1119–25.
33. Claudepierre P, Sibilia J, Goupille P, Flipo RM, Wendling D, Eulry F, et al. Evaluation of a French version of the Bath Ankylosing Spondylitis Disease Activity Index in patients with spondyloarthropathy. J Rheumatol 1997;24:1954–8.
34. Waldner A, Cronstedt H, Stenstrom CH. The Swedish version of the Bath Ankylosing Spondylitis Disease Activity Index: reliability and validity. Scand J Rheumatol Suppl 1999;111:10–6.
35. Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, et al. A comparative study of the usefulness of the Bath Ankylosing Spondylitis Functional Index and the Dougados Functional Index in the assessment of ankylosing spondylitis. J Rheumatol 1999;26:961–5.
36. Akkoc Y, Karatepe AG, Akar S, Kirazli Y, Akkoc N. A Turkish version of the Bath Ankylosing Spondylitis Disease Activity Index: reliability and validity. Rheumatol Int 2005;25:280–4.
37. Brandt J, Westhoff G, Rudwaleit M, Listing J, Zink A, Braun J, et al. Adaption and validation of the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) for use in Germany. Z Rheumatol 2003;62: 264–73. In German.
38. Bonisch A, Ehlebracht-Konig I. The BASDAI-D—an instrument to defining disease status in ankylosing spondylitis and related diseases. Z Rheumatol 2003;62:251–63. In German.
39. El Miedany Y, Youssef S, Mehanna A, Shebrya N, Abu GS, El Gaafaly M. Defining disease status in ankylosing spondylitis: validation and cross-cultural adaptation of the Arabic Bath Ankylosing Spondylitis Functional Index (BASFI), the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), and the Bath Ankylosing Spondylitis Global score (BASG). Clin Rheumatol 2008;27:605–12.
40. Cardiel MH, Londono JD, Gutierrez E, Pacheco-Tena C, Vazquez-Mellado J, Burgos-Vargas R. Translation, cross-cultural adaptation, and validation of the Bath Ankylosing Spondylitis Functional Index (BASFI), the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and the Dougados Functional Index (DFI) in a Spanish speaking population with spondyloarthropathies. Clin Exp Rheumatol 2003;21:451–8.
41. Pimentel-Santos FM, Santos H, Barcelos A, Cunha I, Branco JC, Lopes Ferreira P. Portuguese version of the Bath indices for ankylosing spondylitis patients: a cross-cultural adaptation and validation. Ann Rheum Dis 2010;69 Suppl 3:697.
42. Calin A, Nakache JP, Gueguen A, Zeidler H, Mielants H, Dougados M. Defining disease activity in ankylosing spondylitis: is a combination of variables (Bath Ankylosing Spondylitis Disease Activity Index) an appropriate instrument? Rheumatology (Oxford) 1999;38:878–82.
43. Van Tubergen A, Debats I, Ryser L, Londono J, Burgos-Vargas R, Cardiel MH, et al. Use of a numerical rating scale as an answer modality in ankylosing spondylitis–specific questionnaires. Arthritis Rheum 2002; 47:242–8.
44. Pavy S, Brophy S, Calin A. Establishment of the minimum clinically important difference for the bath ankylosing spondylitis indices: a prospective study. J Rheumatol 2005;32:80–5.
45. Dougados M, Luo MP, Maksymowych WP, Chmiel JJ, Chen N, Wong RL, et al, for the Atlas Study Group. Evaluation of the patient acceptable symptom state as an outcome measure in patients with ankylosing spondylitis: data from a randomized controlled trial. Arthritis Rheum 2008;59:553–60.
46. Tubach F, Pham T, Skomsvoll JF, Mikkelsen K, Bjorneboe O, Ravaud P, et al. Stability of the patient acceptable symptomatic state over time in outcome criteria in ankylosing spondylitis. Arthritis Rheum 2006;55: 960–3.
47. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994;21:2281–5.
48. Stone MA, Pomeroy E, Keat A, Sengupta R, Hickey S, Dieppe P, et al. Assessment of the impact of flares in ankylosing spondylitis disease activity using the Flare Illustration. Rheumatology (Oxford) 2008;47: 1213–8.
49. Boonen A, de Vet H, van der Heijde D, van der Linden S. Work status and its determinants among patients with ankylosing spondylitis: a systematic literature review. J Rheumatol 2001;28:1056–62.
50. Wei JC, Wong RH, Huang JH, Yu CT, Chou CT, Jan MS, et al. Evaluation of internal consistency and re-test reliability of Bath ankylosing spondylitis indices in a large cohort of adult and juvenile spondylitis patients in Taiwan. Clin Rheumatol 2007;26:1685–91.
51. Ruof J, Sangha O, Stucki G. Comparative responsiveness of 3 functional indices in ankylosing spondylitis. J Rheumatol 1999;26:1959–63.
52. Eyres S, Tennant A, Kay L, Waxman R, Helliwell PS. Measuring disability in ankylosing spondylitis: comparison of bath ankylosing spondylitis functional index with revised Leeds Disability Questionnaire. J Rheumatol 2002;29:979–86.
53. Jones SD, Calin A, Steiner A. An update on the Bath Ankylosing Spondylitis Disease Activity and Functional Indices (BASDAI, BASFI): excellent Cronbach's alpha scores. J Rheumatol 1996;23:407.
54. Madsen OR, Rytter A, Hansen LB, Suetta C, Egsmose C. Reproducibility of the Bath Ankylosing Spondylitis Indices of disease activity (BASDAI), functional status (BASFI) and overall well-being (BAS-G) in anti-tumour necrosis factor-treated spondyloarthropathy patients. Clin Rheumatol 2010;29:849–54.
55. Haywood KL, Garratt AM, Dawes PT. Patient-assessed health in ankylosing spondylitis: a structured review. Rheumatology (Oxford) 2005; 44:577–86.
56. Gorman JD, Sack KE, Davis JC Jr. Treatment of ankylosing spondylitis by inhibition of tumor necrosis factor alpha. N Engl J Med 2002;346: 1349–56.
57. Dougados M, Behier JM, Jolchine I, Calin A, van der Heijde D, Olivieri I, et al. Efficacy of celecoxib, a cyclooxygenase 2-specific inhibitor, in the treatment of ankylosing spondylitis: a six-week controlled study with comparison against placebo and against a conventional nonsteroidal antiinflammatory drug. Arthritis Rheum 2001;44:180–5.
58. Heikkila S, Viitanen JV, Kautiainen H, Kauppi M. Does improved spinal mobility correlate with functional changes in spondyloarthropathy after short-term physical therapy? J Rheumatol 2000;27:2942–4.
59. Van Tubergen A, Landewe R, van der Heijde D, Hidding A, Wolter N, Asscher M, et al. Combined spa–exercise therapy is effective in patients with ankylosing spondylitis: a randomized controlled trial. Arthritis Rheum 2001;45:430–8.
60. Jones SD, Steiner A, Garrett SL, Calin A. The Bath Ankylosing Spondylitis Patient Global Score (BAS-G). Br J Rheumatol 1996;35:66–71.
61. Jenkinson TR, Mallorie PA, Whitelock HC, Kennedy LG, Garrett SL, Calin A. Defining spinal mobility in ankylosing spondylitis (AS): the Bath AS Metrology Index. J Rheumatol 1994;21:1694–8.
62. Van der Heijde D, Landewe R, Feldtkeller E. Proposal of a linear definition of the Bath Ankylosing Spondylitis Metrology Index (BASMI) and comparison with the 2-step and 10-step definitions. Ann Rheum Dis 2008;67:489–93.
63. Brandt J, Khariouzov A, Listing J, Haibel H, Sorensen H, Grassnickel L,

et al. Six-month results of a double-blind, placebo-controlled trial of etanercept treatment in patients with active ankylosing spondylitis. Arthritis Rheum 2003;48:1667–75.

64. Machado P, Landewe R, Braun J, Hermann KG, Baker D, van der Heijde D. Both structural damage and inflammation of the spine contribute to impairment of spinal mobility in patients with ankylosing spondylitis. Ann Rheum Dis 2010;69:1465–70.
65. Jones SD, Porter J, Garrett SL, Kennedy LG, Whitelock H, Calin A. A new scoring system for the Bath Ankylosing Spondylitis Metrology Index (BASMI) [letter]. J Rheumatol 1995;22:1609.
66. Heikkila S, Viitanen JV, Kautiainen H, Kauppi M. Functional long-term changes in patients with spondylarthropathy. Clin Rheumatol 2002;21: 119–22.
67. Shinjo SK, Goncalves R, Kowalski S, Goncalves CR. Brazilian-Portuguese version and applicability questionnaire of the mobility index for ankylosing spondylitis. Clinics (Sao Paulo) 2007;62:139–44.
68. Tzelepis GE, Kalliakosta G, Tzioufas AG, Sfikakis PP, Mandros C, Boki KA, et al. Thoracoabdominal motion in ankylosing spondylitis: association with standardised clinical measures and response to therapy. Ann Rheum Dis 2009;68:966–71.
69. Wanders A, Landewe R, Dougados M, Mielants H, van der Linden S, van der Heijde D. Association between radiographic damage of the spine and spinal mobility for individual patients with ankylosing spondylitis: can assessment of spinal mobility be a proxy for radiographic evaluation? Ann Rheum Dis 2005;64:988–94.
70. Jauregui E, Conner-Spady B, Russell AS, Maksymowych WP. Clinimetric evaluation of the bath ankylosing spondylitis metrology index in a controlled trial of pamidronate therapy. J Rheumatol 2004;31:2422–8.
71. Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. J Rheumatol 1988;15:302–7.
72. Landewe R, Dougados M, Mielants H, van der Tempel H, van der Heijde D. Physical function in ankylosing spondylitis is independently determined by both disease activity and radiographic damage of the spine. Ann Rheum Dis 2009;68:863–7.
73. Van der Heijde D, Baraf HS, Ramos-Remus C, Calin A, Weaver AL, Schiff M, et al. Evaluation of the efficacy of etoricoxib in ankylosing spondylitis: results of a fifty-two–week, randomized, controlled study. Arthritis Rheum 2005;52:1205–15.
74. Ruof J, Sangha O, Stucki G. Evaluation of a German version of the Bath Ankylosing Spondylitis Functional Index (BASFI) and Dougados Functional Index (D-FI). Z Rheumatol 1999;58:218–25.
75. Salaffi F, Stancati A, Silvestri A, Carotti M, Grassi W. Validation of the Italian versions of the Bath Ankylosing Spondylitis Functional Index (BASFI) and the Dougados Functional Index (DFI) in patients with ankylosing spondylitis. Reumatismo 2005;57:161–73. In Italian.
76. Karatepe AG, Akkoc Y, Akar S, Kirazli Y, Akkoc N. The Turkish versions of the Bath Ankylosing Spondylitis and Dougados Functional Indices: reliability and validity. Rheumatol Int 2005;25:612–8.
77. Kennedy LG, Edmunds L, Calin A. The natural history of ankylosing spondylitis. Does it burn out? J Rheumatol 1993;20:688–92.
78. Averns HL, Oxtoby J, Taylor HG, Jones PW, Dziedzic K, Dawes PT. Smoking and outcome in ankylosing spondylitis. Scand J Rheumatol 1996;25:138–42.
79. Ward MM, Kuzis S. Validity and sensitivity to change of spondylitis-specific measures of functional disability. J Rheumatol 1999;26:121–7.
80. Daltroy LH, Larson MG, Roberts NW, Liang MH. A modification of the Health Assessment Questionnaire for the spondyloarthropathies. J Rheumatol 1990;17:946–50.
81. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
82. Krishnan E, Sokka T, Hakkinen A, Hubert H, Hannonen P. Normative values for the Health Assessment Questionnaire disability index: benchmarking disability in the general population. Arthritis Rheum 2004;50:953–60.
83. Hidding A, van der Linden S. Factors related to change in global health after group physical therapy in ankylosing spondylitis. Clin Rheumatol 1995;14:347–51.
84. Hidding A, van der Linden S, Gielen X, de Witte L, Dijkmans B, Moolenburgh D. Continuation of group physical therapy is necessary in ankylosing spondylitis: results of a randomized controlled trial. Arthritis Care Res 1994;7:90–6.
85. Viitanen JV, Heikkila S. Functional changes in patients with spondylarthropathy: a controlled trial of the effects of short-term rehabilitation and 3-year follow-up. Rheumatol Int 2001;20:211–4.
86. Queiro R, Sarasqueta C, Belzunegui J, Gonzalez C, Figueroa M, Torre-Alonso JC. Psoriatic spondyloarthropathy: a comparative study between HLA-B27 positive and HLA-B27 negative disease. Semin Arthritis Rheum 2002;31:413–8.
87. Shinjo SK, Goncalves R, Kowalski S, Goncalves CR. Brazilian-Portuguese version of the Health Assessment Questionnaire for Spondyloarthropathies (HAQ-S) in patients with ankylosing spondylitis: a translation, cross-cultural adaptation, and validation. Clin Rheumatol 2007;26:1254–8.
88. Ozcan E, Yilmaz O, Tutoglu A, Bodur H. Validity and reliability of the Turkish version of the Health Assessment Questionnaire for the Spondyloarthropathies. Rheumatol Int. E-pub ahead of print.
89. Moncur C. Ankylosing spondylitis measures. Arthritis Care Res 2003;49 Suppl: S197–209.

Summary Table for Ankylosing Spondylitis Measures*

Scale	Purpose/ content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
ASDAS	Measures disease activity	Self-report	<2 minutes	Hand/computer score	Score from 0 (no disease activity), higher values reflecting higher disease activity	Not reported	Content, construct validity	ES 2.04, SRM 1.45 for improvement with anti-TNF therapy Clinically important improvement 1.1 units, major improvement 2.0 units	Measures important concept with stronger content validity than the BASDAI, extensive validity evidence Appropriate for research use	Still being validated Reliability evidence not reported Use in a clinical setting requiring further assessment
ASQoL	Measures quality of life	Self-report	Median 4 minutes (range 2–16 minutes)	Hand score	0–18, higher scores reflecting greater impairment of quality of life	Internal consistency, test–retest stability	Content, construct validity	mSRM 0.35 (improvement) mSRM 0.57 (deterioration)	Measures important concept, psychometric properties sound, appropriate for clinical and research use	Information on MCID and PASS is lacking
BASDAI	Measures disease activity	Self-report	Mean 67 seconds (range 30 seconds–2 minutes)	Hand score	0 (none or no symptoms) to 10 (very severe symptoms)	Internal consistency, test–retest stability	Content, construct validity	mSRM 0.74 (improvement) mSRM 0.60 (deterioration) ES 1.86, SRM 1.36 for improvement with anti-TNF therapy MCID 10 mm (22.5%) PASS cut off 34.5 mm	Measures important concept, psychometric properties sound, appropriate for clinical and research use	Score based solely on patient-report may omit important objective elements of disease activity
BASFI	Measures functional status	Self-report	<3 minutes	Hand score	0 (no functional impairments) to 10 (maximal impairment)	Internal consistency, test–retest stability, interrater reliability	Content, construct, and criterion validity	ES 0.36, SRM 0.46 (improvement) ES 0.70, SRM 0.72 (deterioration) MCID 7 mm (17.5 %)	Measures important concept, psychometric properties sound, appropriate for clinical and research use	Less sensitive in the well AS patient
BAS-G	Effect of AS on well-being	Self-report	<1 minute	Hand score	0 (no effect on well-being) to 10 (very severe effect on well-being)	Test–retest stability	Construct validity	MCID 15 mm (27.5%)	Measures important concept, psychometric properties sound, appropriate for use in a clinical setting	Less well evaluated than other scales
BASMI	Spinal mobility	Physical measures	5–10 minutes	Hand score	0 (normal spinal mobility) to 10 (severely restricted spinal mobility)	Interrater reliability, Intrarater reliability, test–retest stability	Content, construct, and criterion validity	ES 0.66 (BASMI$_2$) ES 0.95 (BASMI$_{10}$) ES 1.04 (BASMI$_{lin}$)	Measures important concept, appropriate for use in a research setting	Limited by the lack of thoracic spine measures
DFI	Measures functional status	Self-report	Not stated	Hand score	0–40, higher values reflecting higher functional impairment	Interrater reliability, Intrarater reliability, test–retest stability	Content, construct, and criterion validity	ES 0.30, SRM 0.33 (improvement) ES 0.47, SRM 0.59 (deterioration)	Measures important concept, psychometric properties sound, appropriate for research use	5-point Likert scale likely better than original 3-point, but less well validated Information on use in clinical care is lacking
HAQ-S	Measures functional status	Self-report	Not stated	Hand score	0–3, higher values reflect higher impairment	Test–retest stability	Content, construct, and criterion validity	ES 0.20, SRM 0.28 (improvement) ES 0.28, SRM 0.72 (deterioration)	Measures important concept, psychometric properties sound, appropriate for research use	Information on MCID, PASS, and use in clinical care is lacking

* ASDAS = Ankylosing Spondylitis Disease Activity Score; ES = effect size; SRM = standardized response mean; TNF = tumor necrosis factor; BASDAI = Bath Ankylosing Spondylitis Disease Activity Index; ASQoL = Ankylosing Spondylitis Quality of Life scale; MCID = minimum clinically important difference; PASS = patient acceptable symptom state; BASFI = Bath Ankylosing Spondylitis Functional Index; AS = ankylosing spondylitis; BAS-G = Bath Ankylosing Spondylitis Global Score; BASMI = Bath Ankylosing Spondylitis Metrology Index; DFI = Dougados Functional Index; HAQ-S = Health Assessment Questionnaire for the Spondylarthropathies.