Published in final edited form as:

J Clin Epidemiol. 2009 August ; 62(8): 807–815. doi:10.1016/j.jclinepi.2008.10.004.

A Computer-Adaptive Disability Instrument for Lower Extremity Osteoarthritis Research Demonstrated Promising Breadth, Precision and Reliability

Alan M. Jette, PhD, PT¹, Christine M. McDonough, PhD, PT^{1,6}, Stephen M. Haley, PhD, PT¹, Pengsheng Ni, MD, MPH¹, Sippy Olarsch, ScD, PT¹, Nancy Latham, PhD, PT¹, Ronald K. Hambleton, PhD², David Felson, MD, MPH³, Young-jo Kim, MD, PhD⁴, and David Hunter, MD⁵

¹Health & Disability Research Institute, Boston University School of Public Health, Boston, MA

²Center for Educational Assessment, Department of Educational Policy, Research, and Administration, School of Education, University of Massachusetts, Amherst, MA

³Boston University School of Medicine, Boston, MA

⁴Harvard Medical School and Children's Hospital, Boston, MA

⁵New England Baptist Hospital, Boston, MA

⁶The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth College, Hanover, NH

Abstract

Objective—To develop and evaluate a prototype measure (OA-DISABILITY-CAT) for osteoarthritis research using Item Response Theory (IRT) and Computer Adaptive Test (CAT) methodologies.

Study Design and Setting—We constructed an item bank consisting of 33 activities commonly affected by lower extremity (LE) osteoarthritis. A sample of 323 adults with LE osteoarthritis reported their degree of limitation in performing everyday activities and completed the Health Assessment Questionnaire-II (HAQ-II). We used confirmatory factor analyses to assess scale unidimensionality and IRT methods to calibrate the items and examine the fit of the data. Using CAT simulation analyses, we examined the performance of OA-DISABILITY-CATs of different lengths compared to the full item bank and the HAQ-II.

Results—One distinct disability domain was identified. The 10-item OA-DISABILITY-CAT demonstrated a high degree of accuracy compared with the full item bank ($r=0.99$). The item bank and the HAQ-II scales covered a similar estimated scoring range. In terms of reliability, 95% of OA-DISABILITY reliability estimates were over 0.83 versus 0.60 for the HAQ-II. Except at the highest scores the 10-item OA-DISABILITY-CAT demonstrated superior precision to the HAQ-II.

© 2012 Elsevier Inc. All rights reserved.

Corresponding Author: Alan M. Jette, Health & Disability Research Institute, Boston University School of Public Health, 580 Harrison Ave. 2nd floor, Boston, MA 02118. (617) 638-1985, ajette@bu.edu.

Commercial Support/Conflicts Statement. Drs. Haley and Jette have stock interest in CRE Care LLC, which distributes the OA-DISABILITY-CAT Instrument products.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conclusion—The prototype OA-DISABILITY-CAT demonstrated promising measurement properties compared to the HAQ-II, and is recommended for use in LE osteoarthritis research.

Keywords

outcome assessment (Health Care); osteoarthritis; clinical trials; disability; item response theory; computer adaptive testing

Introduction

Disability related to osteoarthritis is widely recognized as a serious problem having significant impact at the individual and societal levels. (1–8) Consequently, disability assessment has become important in the evaluation of persons with osteoarthritis. (9–12) In contrast to functional limitation, which can be defined as a restriction in performance of a specific task at the person level (e.g. put on shoes), disability can be viewed as a limitation in performance of activities within the context of social roles (e.g. visit friends). (13, 14)

Persons with osteoarthritis exhibit a wide range of disability and possess the potential to make large changes over the course of treatment, making it difficult to develop one outcome instrument which works well in all patient groups. (7, 9, 15) An ideal disability measurement instrument would cover the full range of activities relevant to osteoarthritis treatment, with a sufficient number of increments in response categories to measure meaningful change across the disability continuum. Furthermore, to understand the impact of arthritis interventions on the disablement process, measurement instruments would be based on a sound conceptual model, allowing differentiation between functional limitation and disability outcomes.(5, 14, 16–18)

Current disability measures characterize the significant progress in disability assessment made to date, but persistent challenges continue to be faced by developers and users of instruments (9, 11, 19–22). These problems include incomplete coverage of the range of disability levels across patients over the course of treatment, and obtaining adequate precision without excessive instrument length.

A fundamental tension between measurement quality and practicality of administration has persisted for decades. Comprehensive fixed-form instruments have suffered from prohibitive respondent burden and administration costs. (21, 23) The introduction of short-form alternatives has raised concerns over relative losses in score precision and ability to measure clinically meaningful change.(24, 25) These and other well known limitations are largely due to traditional administration methods requiring that a fixed set of questions be administered to all subjects. Often, respondents must address redundant questions or those of low relevance. (26–28) Therefore, while overcoming these stated limitations of existing instruments superior measures of disability would improve the basis for valid judgments about the effectiveness of various osteoarthritis treatments for use in cohort studies.

Contemporary methods for outcome instrument construction and data collection provide an opportunity to improve psychometric properties while reducing respondent burden and administrative costs. The introduction of Item Response Theory (IRT) methods, (29–31) has allowed researchers to develop outcome instruments with improved performance across a broad range of disability. However, IRT methods by themselves have not resolved the problem of respondent burden. Recent introduction of computer adaptive testing (CAT) methods combined with IRT methods in the health measurement field offers the potential solution to this challenge. (32, 33) In CAT administration, an iterative computer program uses information from a subject's previous responses to tailor item selection to provide the most information at the respondent's current score estimate, thereby eliminating redundant

questions about activities that are too hard or too easy. A key strength of this approach is that all scores are on the same metric, regardless of the number of items administered, thus facilitating comparisons across time or across groups with different disability levels.(34)

In this study we developed a disability instrument for lower extremity (LE) osteoarthritis research (OA-DISABILITY-CAT) using IRT and CAT methodologies, and evaluated its psychometric performance in relation to the full item bank and to the Assessment Questionnaire Disability Index (HAQ-II), a widely used health measure in the field.

Methods

Instruments

HAQ—The Health Assessment Questionnaire (20) is a generic, multi-dimensional instrument designed to measure function and disability outcomes for rheumatoid arthritis (RA) that has been commonly used in many other disease areas, including osteoarthritis.(19) Subsequent versions of the HAQ including the MHAQ, MDHAQ, and the HAQ-II have addressed limitations of the original HAQ. In this study, we used the 10-item HAQ-II (22, 35). In a comparison of the HAQ-II with the earlier versions of the HAQ, and the SF-36, the HAQ-II demonstrated a reliability estimate of 0.88; was highly correlated with the other measures, and had fewer ceiling and floor effects than the earlier HAQs. (22) Higher scores indicate more disability for the HAQ-II; however for consistency in this study we reversed scores so that higher scores indicate less disability.

OA-DISABILITY-CAT Item Bank Development

Patient Focus Groups—We conducted 6 semi-structured focus groups each consisting of 5 to 6 patients with LE osteoarthritis. Experienced moderators elicited patients' perspectives on important outcomes for osteoarthritis research. Transcripts of audiotapes of the sessions were content analyzed.

Clinician Focus Groups—We directed 3 multi-disciplinary focus groups including 5 to 6 clinicians who had extensive expertise in the treatment of patients with osteoarthritis.

Literature Review—We conducted a comprehensive review of the literature and generated a list of daily life activities covering a broad range of disability levels. The final item bank consisted of 33 daily life activities commonly affected by LE osteoarthritis.

Cognitive Testing—We performed cognitive testing on the item bank to identify problems with questions that would diminish instrument performance by asking 6 adult patients with LE osteoarthritis scripted questions about item meaning.

In the final scale, subjects were asked to report the amount of limitation they had doing each activity as: 1) Not at all limited 2) A little, 3) A lot, 4) Did not do this activity because of the arthritis in my legs, 5) Did not do an activity for reasons other than the arthritis in my legs. The time frame "on an average day over the past month" was used. If the subject responded, 'did not do activity because of arthritis', those responses were treated as 'couldn't do activity'; if the subject responded, 'did not do an activity for reasons other than arthritis', those responses were treated as missing data for the analysis.

Study Sample—We recruited a convenience sample of 323 adults from the greater Boston area from a pool of patients who had previously participated in osteoarthritis research and from a local orthopedic surgeon's practice. In all cases the diagnosis of knee and/or hip osteoarthritis was confirmed by a physician, and the patient experienced pain or stiffness

within the past 30 days consistent with the ACR clinical criteria for defining osteoarthritis. For the majority of the sample, arthritis was confirmed by radiographic evidence as well.

Data Collection—Subjects were contacted by phone to determine eligibility which included: 18 years or older, able to speak English, experienced pain or stiffness in their knee or hip within the prior month, evidence on radiograph of a definite osteophyte for the knee or hip or joint space narrowing for the hip or confirmation from the subject of a physician's diagnosis of osteoarthritis of the knee or hip. Subjects were not eligible if they used a wheelchair in their home, or had been diagnosed with rheumatoid arthritis, systemic lupus erythematosus, gout, or psoriatic arthritis. Subjects were stratified by functional level, ascertained by the Physical Function domain of the SF-36 to ensure a range of functional ability in the sample.

The OA-DISABILITY-CAT item bank and HAQ-II items were administered by trained interviewers during a home visit with each subject. The HAQ-II was administered first using pen and paper, followed by computerized administration of three instruments, including the OA-DISABILITY-CAT. The order of the three computerized instruments was counterbalanced. Gender-specific items were administered to the relevant gender. Demographic information (age, sex, ethnicity, race, education, living and housing status) were collected for each subject. All procedures were approved by the Institutional Review Board at Boston University.

OA-DISABILITY-CAT Structure/Unidimensionality—We tested the underlying structure of the proposed disability items in a series of confirmatory factor analyses (36) and evaluated item loadings and residual correlations between items using MPlus software.(37) To maximize precision in our evaluation of these skewed categorical data, we chose unweighted least squares (ULS) estimation based on polychoric correlation matrices and variance adjusted estimation methods. (36, 38) We assessed eigenvalues associated with each factor extracted. Our analysis of model fit included the ratio of chi-square to degrees of freedom, Comparative Fit Index, (CFI) Tucker-Lewis Index, (TLI) and Root Mean Square Error Approximation (RMSEA). For CFI and TLI values range from 0 to 1, with higher values indicating better test model fit compared to a baseline model, and 0.95 or greater representing acceptable fit. RMSEA represents misfit per degree of freedom, and lower values signify better fit. Values less than 0.05 suggest a “very good fit”, with values around 0.08 interpreted as “mediocre” fit. Values >0.1 are generally viewed as indicative of a “poor fit.” (39, 40) Our second approach used the magnitude of the factor loadings on the primary factor. Finally, we considered residual correlations; those less than or equal to 0.20 suggest that the primary factor explains the correlation between items, and indicates acceptable fit. (41, 42) Higher residual correlations signified violation of the local independence assumption.

Item Calibrations—The item calibrations were estimated using the generalized partial credit model (GPCM).(43–45) We estimated IRT-based scores for the disability domain using Weighted Maximum Likelihood Estimation. (38, 46) We evaluated fit using the likelihood ratio chi-square (G^2) statistics for each item based on the comparison of expected and observed values across the distribution of the domain. Bonferroni corrected p-values were used in the significance tests and the likelihood ratio chi-square statistic for the whole test was also examined to verify model fit of the domain. The scores estimated from the IRT model were standardized to have a mean of 50 and standard deviation of 10. All of the IRT analyses was performed using the software package PARSCALE. (47)

Differential Item Functioning—A basic assumption of IRT models is that a subject's score on an item should depend entirely on the subject's score in the domain being measured

and the statistical characteristics of the item. Significant differential item function (DIF) indicates that background variables (such as age or gender) influenced the response. (48) There are two kinds of DIF; uniform DIF, which means the item response difference was constant across the reporting scale for ability; and non-uniform DIF, which means the item response difference between, say, males and females, was not consistent across the score reporting scale for disability.

DIF was assessed using logistic regression, with the OA-DISABILITY-CAT item score chosen as the dependent variable and background variables assigned as the independent variable. In the DIF analysis if the background effect was significant and the interaction effect with a person's disability level was not, then the item had uniform DIF; on the other hand, if the interaction effect was significant, the item had non-uniform DIF. The analytic strategy successively added disability levels, background variables and interaction terms into the model and model comparison was based on the likelihood ratio test. The effect size of the DIF was classified based on the R-square change between models. (49)

Development of the Simulated CAT Program

Once a final item bank was identified and item calibrations were generated for the disability domain, we constructed the OA-DISABILITY-CAT algorithms on HDRI™ software developed at Boston University. The CATs were designed to be administered from a stand-alone computer or from a web-based platform. We programmed the CATs to use weighted maximum likelihood (WML) score estimation and selected initial items from those in the middle of the pain and disability ranges. The response to the first item was fed into the CAT algorithm and the application calculated a probable score as well a person-specific standard error (measure of precision). Additional questions were selected and administered until the maximum number of items had been administered (in our analyses, 5, 10 or 15 items were administered).

An assumption of IRT is that all items are locally independent, that is, patients' responses to any pair of items are statistically independent. (29) Often, items with local dependence are removed from the item bank. In our case, we did not eliminate them from the item bank, but rather dealt with them by special programming within the CAT algorithm that allowed use of only one item within a set of locally dependent items.

Psychometric Evaluation of the OA-DISABILITY-CAT

We conducted simulations to estimate the performance of CATs of different fixed item lengths (i.e. 5-, 10-, and 15 items) with respect to the full item bank. Mean scores generated by CATs of different lengths were compared with scores generated by the full item bank for the entire sample and across osteoarthritis conditions. To compare the relative precision of the CAT scores at multiple points along the scale with the full item bank we plotted the standard errors in relation to each subject's disability scores. Pearson correlations were calculated between each of the CAT-generated scores and the full item bank scores to estimate the CAT's accuracy. We compared the OA-DISABILITY-CAT item distributions, floor and ceiling effects, reliability, and precision against the HAQ-II. To create an appropriate comparison we placed the HAQ-II and OA-DISABILITY-CAT scores on the same metric by fixing the calibrations of one of the instruments and placing the other one on this same scale. Thus, we calibrated the HAQ-II items by anchoring the OA-DISABILITY-CAT item calibrations in the disability domain. For both scales in this analysis, higher scores indicate less disability. To examine the distribution of the DISABILITY-CAT and HAQ-II items, we calculated expected values for each response category for each item. We considered the range of the scale to be the corresponding person score estimates between the expected value of the lowest and highest response category in each scale. In addition, we

calculated the percent at the ceiling and floor for each scale. We compared the relative precision of the OA-DISABILITY-CAT scores with the HAQ-II using standard errors. Reliability, the degree to which the differences across patient measurements are due to actual differences in disability (true variance) rather than to measurement error, was examined by comparing the ratio of the true variance to the total variance for each instrument at multiple points along the scale. Reliability was estimated as follows: $1/(1 + (\text{standard error})^2)$. (50) Any section of the reliability function <0.70 was considered to be inadequate. To test construct validity of the OA-DISABILITY instruments, we calculated Pearson correlation coefficients between the HAQ-II and the OA-DISABILITY instruments (5-, 10-, 15- item CATs and the full item bank). We hypothesized that the correlations would be strong (>0.60).

Results

In the study sample, the average age was 62 years, (sd = 15), 65% of participants were female, and a large proportion had knee osteoarthritis (Table 1). In the disability scale the average percentage of subjects who responded “didn’t do an activity because of arthritis” was 9.21% (sd 12.26%). The average percentage of subjects who responded “didn’t do an activity for reasons other than arthritis” was 18.18%, (sd 13.34%).

DISABILITY Domain

Confirmatory factor analysis results were consistent with unidimensionality of the OA-DISABILITY-CAT domain. A unidimensional model (chi-square(df)=251(96), $p<0.0001$) across all 33 items achieved an acceptable level of fit, explained 62% of the variance, and was easily interpretable. Only 1.4% of the residual covariances were greater than ± 0.20 , which means that the local independence assumption was satisfied. Remaining fit statistics were as follows: CFI was 0.95; TLI was 0.99; and RMSEA was 0.07.

The data fit the generalized partial credit model; the chi-square (df)=320(338), $p=0.76$. In terms of item fit, there was only 1 misfitting item (fairly heavy house or yard work) in the item bank.

Differential item functioning (DIF)

There were 3 items which displayed DIF by age (taking part in a regular exercise program, doing low demand sports such as golfing or bowling and using public transportation). After adjusting for disability level, taking part in a regular exercise program was more difficult for those with hip osteoarthritis, and therefore demonstrated DIF by osteoarthritis condition. No items displayed gender DIF, and only uniform DIF was detected in this analysis.

Comparison of OA-DISABILITY-CAT to Full Item Bank

Pearson correlation coefficients between the 5-, 10, and 15-item OA-DISABILITY-CATs and the full item banks were 0.93, 0.99, and 0.99 respectively. This high degree of accuracy is illustrated by score plots for the 10-item CAT and the full item bank (Figure 1). Table 2 shows that the descriptive statistics of scores from the 5- 10- and 15-item OA-DISABILITY-CATs were similar to those for the full item bank and for mean scores generated across osteoarthritis conditions. As might be expected, the 5-item CAT had a smaller range of scores than the 10-, 15, and the full item bank. The standard errors of the 10-item OA-DISABILITY-CAT were slightly larger than the full item bank scores across the range reflecting the fewer number of items that were used to calculate the overall score.

Comparison with the HAQ-II

In Figure 2 the breadth of item and response category coverage across the continuum for the OA-DISABILITY item bank is displayed relative to that of the HAQ-II scale. Item and response category coverage is displayed as the range of scores for the sample that correspond to the highest and lowest values of expected item response categories in each scale. The OA-DISABILITY item bank and the HAQ-II scales covered a similar estimated scoring range (Figure 2). The ceiling and floor calculations further illustrated these results, for example, 13 (4.02%) of subjects were at the ceiling (scores indicating least disability) for the OA-DISABILITY item bank compared to 18 (5.57%) for the HAQ-II scale. No floor effects (scores indicating most disability) were detected for either scale.

Correlations between the HAQ-II and the OA-DISABILITY-CATs (5-, 10-, 15- item CATs and the full item bank) ranged from 0.71 to 0.74.

The conditional reliability of the OA-DISABILITY item bank was very strong across the center of the disability continuum. For example, 95% of OA-DISABILITY reliability estimates were over 0.83 versus 0.60 for the HAQ-II (Figure 3). Reliabilities decreased as the level of disability approached the ceiling for both the OA-DISABILITY item bank and the HAQ-II; however, the OA-DISABILITY item bank reliability remained superior to the HAQ-II. At the extreme floor, reliability for the HAQ-II surpassed that of the OA-DISABILITY-CAT.

Figure 4 displays the precision of the 10-item OA-DISABILITY-CAT and the HAQ-II scales as measured by the conditional standard error of measurement statistic. The 10-item OA-DISABILITY-CAT demonstrated superior precision to the HAQ-II for much of the range of scores; however, at the very highest scores precision of the two instruments was similar, and in some cases the precision of the HAQ-II exceeded that of the OA-DISABILITY-CAT.

Discussion

The results of these analyses revealed that the OA-DISABILITY item bank and the OA-DISABILITY-CAT scales performed well in this sample of persons with LE osteoarthritis. The full 33 item bank calibrated well with a unidimensional IRT model, providing similar breadth and, on average, more precise and reliable estimates of disability than the HAQ-II. The OA-DISABILITY item bank was similar to the HAQ-II with respect to breadth of coverage of the continuum of disability while the 10-item OA-DISABILITY-CAT showed improved reliability and precision throughout much (but not all) of the disability continuum as compared with the HAQ-II. Therefore, the OA-DISABILITY-CAT will be of particular benefit in providing precise and reliable measurement of disability with few items. Nonetheless, further improvements could be made to this scale at the ceiling.

The correlations found between the OA-DISABILITY instruments and the HAQ-II were very strong, (51, 52) and offer one indication that the OA-DISABILITY-CAT item bank provides a valid representation of the repercussions of LE osteoarthritis. In spite of the fact that the two scales have important differences, we would expect that there would be a trend toward increased OA-DISABILITY-CAT scores as HAQ-II scores increase.

In previous research, the HAQ-II has demonstrated strengths in breadth of coverage of the severity of the impact of arthritis. However, it includes items that measure both function (e.g. lift heavy objects) and disability (e.g. do outside work) within the same instrument. This poses a major barrier to investigation of the disablement process, where it is critically important to use distinct measures of impairments, functional limitations, and disability.(17)

The incremental improvements of the OA-DISABILITY-CAT over the HAQ-II in terms of reliability and measurement precision are augmented by the considerable advantage of focusing specifically on disability. Instruments such as the OA-DISABILITY-CAT and the OA-FUNCTION-CAT (53) which strive for conceptual clarity allow research to uncover factors that impact the development or resolution of disability given that impairment and/or functional limitations have occurred.(14, 17)

Although preliminary, the results from the present study are encouraging and consistent with previous work indicating that the 10- item CATs will likely provide the opportunity to maximize psychometric properties with minimal data collection time and administrative burden. (34, 35) Further work is needed to ascertain the administrative burden and responsiveness to clinically meaningful change.

At this stage of the development of the item banks, we did not remove any items due to DIF. However, there were some interesting results. Taking part in a regular exercise program, doing low demand sports such as golfing or bowling and using public transportation were more difficult for older patients. Taking part in a regular exercise program was more difficult for those with hip osteoarthritis. These predictable patterns of differences support construct validity of our instrument.

DIF can be handled in several ways. One approach is to simply remove the items from the calibrated item bank and only use those without DIF. One disadvantage of this approach is that it may diminish the sensitivity of the resulting item banks and thus reduce the utility of the CAT instrument. An alternative approach would be to establish different sets of calibrations for hip and knee patients and incorporate them into future CAT applications. We are especially interested in pursuing this second approach in future research.

Several limitations of the research should be acknowledged. Although the sample used in this study was adequate, it was relatively small for an IRT analysis. One consequence is that the person and item standard errors were larger than might be desirable for wider uses of the item banks. Secondly, the effect of sample size on the number of unexpected responses for any particular item in the bank could potentially have lead to erroneously labeling an item “fitting” when with a larger sample; the opposite evaluation may be true. Finally, the impact of sample size for DIF analysis is that our results could have underestimated the presence of DIF. Clearly the structure of the OA-DISABILITY-CAT revealed in this study needs to be replicated in other samples with LE osteoarthritis.

In addition, real data simulations are based on the assumption that the answers to a subset of those items selected using CAT would be identical to the answers given if they were embedded in a larger fixed-form instrument, such as was administered to the calibration sample. Such simulations are likely good (but not perfect) approximations of actual CAT administrations and may overestimate the score agreement of CATs with the full item bank. Future research needs to examine the accuracy of CAT estimates in prospective studies.

Conclusions

This study revealed that the OA-DISABILITY item bank and 10-item OA-DISABILITY-CAT provided comparable or superior measurement properties compared to a widely used traditional measure in a sample of patients with LE osteoarthritis. The strong conceptual basis for this disability scale, combined with incremental improvements in reliability and precision compared to the HAQ-II support OA-DISABILITY-CAT as a strong candidate for future measurement of osteoarthritis-related disability. Further work is needed to test the performance of the OA-DISABILITY-CAT prospectively. This preliminary study and the evolving body of work indicate that the CAT approach combined with IRT offers a viable

solution to the longstanding conflict between the need for accuracy in clinical assessment and the equal need for practicality of administration.

Acknowledgments

Supported by the NIH R01 AR 051870 and 1F32HD056763 and an Independent Scientist Award (K02 HD45354-01) to Dr. Haley

References

1. Covinsky K. Aging, Arthritis, and Disability. *Arthritis & Rheumatism (Arthritis Care & Research)*. 2006 Apr 15; 55(2):175–6. [PubMed: 16583390]
2. Dunlop D, Hughes S, Manheim L. Disability in activities of daily living: Patterns of change and a hierarchy of disability. *Am J Public Health* 1997. 1997; 87:378–383.
3. Dunlop DD, Manheim LM, Song J, Chang RW, Dunlop DD, Manheim LM, et al. Arthritis prevalence and activity limitations in older adults. *Arthritis Rheum*. 2001 Jan; 44(1):212–21. [PubMed: 11212163]
4. Guralnik J, Simonsick E, Ferrucci L, 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. 1994; 49:M85–94.
5. Institute of Medicine (IOM). *The Future of Disability in America*. Washington, DC: The National Academies Press; 2007.
6. Song J, Chang RW, Dunlop DD. Population Impact of Arthritis on Disability in Older Adults. *Arthritis & Rheumatism (Arthritis Care & Research)*. Apr 15; 2006 55(2):248–55. [PubMed: 16583415]
7. Verbrugge LM, Juarez L. Profile of Arthritis Disability: II. *Arthritis & Rheumatism (Arthritis Care & Research)*. 2006 Feb 15; 55(1):102–13. [PubMed: 16463411]
8. Ling SM, Xue QL, Simonsick EM, Tian J, Bandeen-Roche K, Fried LP, et al. Transitions to Mobility Difficulty Associated With Lower Extremity Osteoarthritis in High Functioning Older Women: Longitudinal Data From the Women's Health and Aging Study II. *Arthritis & Rheumatism (Arthritis Care & Research)*. 2006 Apr 15; 55(2):256–63. [PubMed: 16583416]
9. Veenhof C, Bijlsma JW, Van den Ende CHM, Van Dijk GM, Pisters MF, Dekker J. Psychometric Evaluation of Osteoarthritis Questionnaires: A Systematic Review of the Literature. *Arthritis & Rheumatism (Arthritis Care & Research)*. 2006 Jun 15; 55(3):480–92. [PubMed: 16739188]
10. Bijlsma JW, Bijlsma JJJ. Patient centred outcomes in osteoarthritis. *Ann Rheum Dis*. 2005 Jan; 64(1):1–2. [PubMed: 15608297]
11. Rat AC, Coste J, Pouchot J, Baumann M, Spitz E, Retel-Rude N, et al. OAKHQOL: a new instrument to measure quality of life in knee and hip osteoarthritis. *Journal of Clinical Epidemiology*. 2005 Jan; 58(1):47–55. [PubMed: 15649670]
12. 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 Jan; 64(1):29–33. [PubMed: 15208174]
13. Jette AM. Disablement Outcomes in Geriatric Rehabilitation. *Med Care*. 1997; 35(6):JS28–JS37. [PubMed: 9191712]
14. Jette AM, Keysor JJ. Disability models: Implications for arthritis exercise and physical activity interventions. *Arthritis Care and Research*. 2003 Feb; 49(1):114–20. [PubMed: 12579601]
15. Garratt AM, Brealey S, Gillespie WJ, Team DT, Garratt AM, Brealey S, et al. Patient-assessed health instruments for the knee: a structured review. *Rheumatology (Oxford)*. 2004 Nov; 43(11):1414–23. [PubMed: 15316121]
16. Jette AM. Assessing Disability in Studies on Physical Activity. *American Journal of Preventive Medicine*. 2003; 25(3Sii):122–8. [PubMed: 14552935]
17. Verbrugge LM, Jette AM. The Disablement Process. *Social Science and Medicine*. 1994; 38(1):1–14. [PubMed: 8146699]

18. Jette AM, Haley SM, Kooyoomjian JT, Jette AM, Haley SM, Kooyoomjian JT. Are the ICF Activity and Participation dimensions distinct? *Journal of Rehabilitation Medicine*. 2003 May; 35(3):145–9. [PubMed: 12809198]
19. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, progress, and documentation. *Journal of Rheumatology*. 2003; 30:167–78. [PubMed: 12508408]
20. Fries J, Spitz P, Kraines R, Holman H. Measurement of patient outcome in arthritis. *Arthritis Rheum*. 1980; 23:137–145. [PubMed: 7362664]
21. Ware JE Jr. Conceptualization and measurement of health-related quality of life: comments on an evolving field. *Arch Phys Med Rehabil*. 2003; 84(Supplement 2):S43–51. [PubMed: 12692771]
22. Wolfe F, Michaud K, Pincus T, Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire.[see comment]. *Arthritis Rheum*. 2004 Oct; 50(10):3296–305. [PubMed: 15476213]
23. McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Ann Intern Med*. 1997; 127:743–50. [PubMed: 9382391]
24. Backman CL. Outcome Measures for Arthritis Care Research: Recommendations from the CARE III Conference. *The Journal of Rheumatology*. 2006; 33(9):1908–11. [PubMed: 16960954]
25. Rubenach S, Shadbolt B, McCallum J. Assessing health-related quality of life following myocardial infarction: Is the SF-12 useful? *J Clin Epidemiol*. 2002; 55:306–9. [PubMed: 11864802]
26. Beaton D, Richards R. Measuring Function of the Shoulder. *J Bone Joint Surg*. 1996; 78(A6):882–90. [PubMed: 8666606]
27. Chen A-T, Broadhead W, Doe E, Broyles W. Patient acceptance of two health status measures: The Medical Outcomes Study Short-Form General Health Survey and the Duke Health Profile. *Fam Med*. 1993; 25:536–9. [PubMed: 8405803]
28. McHorney C, Bricker DJ. A qualitative study of patients' and physicians' views about practice-based functional health assessment. *Med Care*. 2002; 40:1113–25. [PubMed: 12409856]
29. Hambleton, RK.; Swaminathan, H.; Rogers, HJ. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications; 1991.
30. Lord, FM. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum Associates; 1990.
31. van der Linden, WJ.; Hambleton, RK. *Handbook of Modern Item Response Theory*. New York: Springer-Verlag New York, Inc; 1997.
32. Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computer adaptive assessment. *Quality of Life Research*. 2007; 16:133–41. [PubMed: 17401637]
33. Jette AM, Haley SM. Contemporary measurement techniques for rehabilitation outcomes assessment. *Journal of Rehabilitation Medicine*. 2005; 37:339–45. [PubMed: 16287664]
34. Wainer, H. *Computer Adaptive Testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
35. Wolfe F, Wolfe F. Why the HAQ-II can be an effective substitute for the HAQ. *Clinical & Experimental Rheumatology*. 2005 Sep-Oct;23(5 Suppl 39):S29–30. [PubMed: 16273782]
36. Mislevy RJ. Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*. 1986; 11:3–31.
37. Muthen, B.; Muthen, L. *MPlus User's Guide*. Los Angeles, CA: Muthen & Muthen; 1998.
38. Beauducel A, Herzberg PY. On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*. 2006; 13:186–203.
39. Browne, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, KA.; Long, JS., editors. *Testing structural equation models*. Thousand Oaks, CA: Sage; 1993. p. 136-62.
40. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*. 1993; 6(1):1–55.

41. Reeve BB, Hays RD, Bjorner JB. 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 May; 45(5 Suppl 1):S22–31. [PubMed: 17443115]
42. Yen WM. Scaling Performance Assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*. 1993; 30:187–213.
43. Andrich, D. Rasch Models for measurement. Beverly Hills, CA: Sage Publications; 1998.
44. Fischer, G.; Molenaar, I. Rasch Models: Foundations, recent developments, and applications. Berlin: Springer-Verlag; 1995.
45. Masters GN. A Rasch model for partial credit scoring. *Psychometrika*. 1982; 47:147–74.
46. Warm TA. Weighted likelihood estimation of ability in item response theory. *Psychometrika*. 1989; 54:427–50.
47. Muraki, E.; RDB. PARSCALE: IRT Item Analysis and Test Scoring for Rating--Scale Data. Chicago: Scientific Software International; 1997.
48. Hariharan S, Rogers HJ. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*. 1990; 27:361–70.
49. Jodoin M, Gierl M. Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*. 2001; 14 (4):329–49.
50. Mâsse LC, Heesch KC, Eason KE, Wilson M. Evaluating the properties of a stage-specific self-efficacy scale for physical activity using classical test theory, confirmatory factor analysis and item response modeling. *Health Educ Res*. 2006; 21:i33–i46. [PubMed: 17035352]
51. McDowell, I. *Measuring Health: A Guide to Rating Scales and Questionnaires*. 3. New York: Oxford University Press; 2006.
52. Streiner, DL.; Norman, GR. *Health Measurement Scales*. New York: Oxford University Press; 2003.
53. Jette, AM.; McDonough, CM.; Ni, P.; Haley, SM.; Hambleton, RK.; Olarsch, S., et al. A Functional Difficulty and Functional Pain Instrument for Lower Extremity Osteoarthritis Research. In Press

Appendix A. OA–DISABILITY-CAT Items

Physical Disability

In this section I will ask you about everyday activities you may have done over the past month. I will ask you to what degree you felt limited in doing each activity because of the arthritis in your legs.

For each activity, please choose from the following answers in describing how limited you felt, on an average day, during the past month, because of the arthritis in your legs:

Not at all limited (1)

A little (2)

A lot (3)

For those activities that you did not do, I want to know if that is because of the arthritis in your legs. (4)

If you did not do an activity for reasons unrelated to arthritis, select the response: “Did not do the activity for reasons other than the arthritis in my legs”. (5)

Because of the arthritis in your legs, how limited did you feel on an average day, over the past month when you were.....?

1. Visiting friends and relatives
2. Traveling out of town for work or vacation
3. Going out with others to public places
4. Working at a volunteer job
5. Taking part in organized social activities
6. Providing care or assistance to others
7. Taking care of the inside of your home
8. Providing meals for yourself and/or your family
9. Taking care of your own personal care needs
10. Taking care of local errands
11. Taking part in a regular exercise program
12. Taking care of children or grandchildren
13. Using public transportation
14. Driving a car
15. Doing vigorous activities such as participating in strenuous sports
16. Having usual sexual activity
17. Going shopping
18. Falling asleep at night
19. Doing your usual lifestyle activities
20. Going to church or temple
21. Going to the mall
22. Working in the garden
23. Playing with children
24. Going to a movie theater
25. Doing fairly heavy house or yard work
26. Doing heavy housework or repairs (e.g., washing windows, shoveling snow)
27. Going to a restaurant for a meal
28. Traveling by airplane
29. Doing your usual paid work
30. Participating in your usual sport or recreational activities
31. Doing high demand sports such as football, basketball, tennis, or aerobic dancing
32. Doing low demand sports such as golfing or bowling
33. In your ability to do your desired sport as long as you would like.

Appendix B. Health Assessment Questionnaire-II (HAQ-II) Items

We are interested in learning how your illness affects your ability to function in daily life. Place an X in the box which best describes your usual abilities over the past week.

Without any difficulty (0)

With some difficulty (1)

With much difficulty (2)

Unable (3)

Are you able to

1. Get on and off the toilet?
2. Open car doors?
3. Stand up from a straight chair?
4. Walk outdoors on flat ground?
5. Wait in a line for 15 minutes?
6. Reach and get down a 5-pound object (such as a bag of sugar) from just above your head?
7. Go up 2 or more flights of stairs?
8. Do outside work (such as yard work)?
9. Lift heavy objects?
10. Move heavy objects?

What is new?

- We report on the development of a new measure of disability for osteoarthritis research that provides superior:
 - Conceptual Clarity
 - Precision
 - Reliability
- This contemporary instrument, developed using Item Response Theory and Computer Adaptive Testing methods, offers a viable solution to the longstanding conflict between the need for accuracy in clinical outcome assessment and for practicality of administration.

Figure 1.
Correlation between the 10-item OA-DISABILITY-CAT Scale scores with the full Item Bank

Figure 2.
Comparison of the breadth of the OA-DISABILITY Item Bank compared with the HAQ-II displayed as the range of the highest and lowest expected item categories in each scale.

Figure 3. Comparison of the OA-DISABILITY Item Bank and HAQ-II reliability estimates across the continuum of disability. Higher scores indicate less disability.

Figure 4. Precision of the 10-item OA-DISABILITY-CAT & HAQ-II as measured by standard errors of scale scores.

Table 1

Characteristics of the study sample

	Total Sample N=323
Mean Age (SD)	61.8 (15.2)
Female N (%)	209 (64.7)
Race/Ethnicity N (%)	
American Indian or Alaskan Native	1 (0.3)
Asian or Pacific Islander	8 (2.5)
Black (Not of Hispanic Origin)	29 (9.0)
Hispanic	4 (1.2)
Mixed Race or Ethnicity	2 (0.6)
Other or Not Available	4 (1.2)
White (Not of Hispanic Origin)	275 (85.1)
Joints Affected: N (%)	
Hip and Knee	88(27.2)
Hip only	53(16.4)
Knee only	182(56.4)

Table 2

Comparison of scores from the 5- 10- and 15-item OA-DISABILITY-CATs and the full Item Bank (N=323)

	Total		Both Hip and Knee		Hip		Knee	
	Mean(SD)	Range	Mean(SD)	Range	Mean(SD)	Range	Mean(SD)	Range
5-CAT	49.65(10.00)	25.94–77.06	47.11(9.02)	26.92–77.06	49.67(8.81)	33.88–66.06	50.89(10.59)	25.94–77.06
10-CAT	50.25(11.08)	27.75–91.49	47.45(9.72)	27.90–76.80	50.49(9.23)	35.03–75.95	51.55(11.97)	27.75–91.49
15-CAT	50.38(11.26)	27.20–92.49	47.56(9.85)	27.20–77.10	50.52(9.11)	33.93–76.21	51.72(12.23)	27.54–92.49
Full Item Bank	50.43(11.23)	26.46–92.48	47.8(9.97)	26.46–77.54	50.40(9.30)	33.63–76.32	51.73(12.12)	26.92–92.48

Higher scores indicate less disability