Clin Orthop Relat Res (2013) 471:3489–3495
DOI 10.1007/s11999-013-3028-1

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

SYMPOSIUM: ABJS CARL T. BRIGHTON WORKSHOP ON OUTCOME MEASURES

Statistical Considerations in the Psychometric Validation of Outcome Measures

Alla Sikorskii PhD, Philip C. Noble PhD

Published online: 4 May 2013

Abstract

Background The evaluation of the outcomes of total knee arthroplasty requires measurement tools that are valid, reliable, and responsive to change. However, the accuracy of any outcome measurement is determined by the validity and reliability of the instrument used. To ensure this accuracy, it is imperative that each instrument used in orthopaedics is free of biases leading to inaccurate estimates of treatment effects.

Where are we now? Many patient-derived outcome instruments have been developed and tested through the application of the standard assessments that form the basis of classical test theory: validity, reliability, and responsiveness. These assessments determine if the instrument reliably measures what it is intended to measure, and if it captures differences among groups of patients or changes over time.

Where do we need to go? Thorough evaluation of the outcome instruments used in orthopaedics is a critical prerequisite for the continued improvement of effective patient care. Additional steps of psychometric testing that are sometimes overlooked include testing for differential item functioning (DIF) and the effects of the mode of administration of the outcome instrument. The use of suitable approaches to test for these potential sources of bias would facilitate the development of more robust outcome assessment in research and clinical practice.

How do we get there? Testing for DIF, including the effects of mode of administration, may be performed using several analytical approaches. This will allow optimal application of each outcome instrument with respect to patient characteristics, time and mode of the administration, and modification, as necessary.

Introduction

The evaluation of patient outcomes after TKA requires measurement tools that are valid, reliable, and responsive to change. Many different measures have been advocated to gauge the effectiveness of TKA and to assess variations in clinical pathways, patient selection, surgical approaches, and different implant designs [3, 6, 9, 21, 22, 30, 37, 39, 48 and references therein]. Because these outcome measurements inform fundamental decisions concerning the efficacy and quality of patient care, it is imperative that these tools be free of biases that could lead to inaccurate estimates of treatment effects. These outcome measures should also have good responsiveness because small estimates of the effects may be the result of lack of the responsiveness in the measures themselves. The lack of effect may be the result of failure of the measure to capture

Each author certifies that he or she, or a member of his or her immediate family, has no funding or commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research* editors and board members are on file with the publication and can be viewed on request.
This work was performed at Michigan State University, East Lansing, MI, USA, and the Institute of Orthopedic Research & Education, Houston, Texas, USA.

A. Sikorskii (✉)
Department of Statistics and Probability,
Michigan State University, 619 Red Cedar Road,
C423 Wells Hall, East Lansing, MI 48824, USA
e-mail: sikorska@stt.msu.edu

P. C. Noble
Barnhart Department of Orthopedic Surgery,
Baylor College of Medicine, Houston, TX, USA

a treatment effect that truly exists or be indicative of the limited treatment itself. This distinction cannot be made without investigating the psychometric properties of the measurement tools; however, few tools, especially those that are patient-reported, have been subjected to comprehensive evaluation [9, 48]. The steps that have been completed to date for various tools vary, and further testing is necessary to inform both clinicians and researchers in selecting the most appropriate instrument for a particular purpose.

The fundamental issue in outcome measurement is that a characteristic of interest (eg, pain, satisfaction), also referred to as a trait or a construct, is not observed and cannot be measured directly. Thus, a responder must be presented with a set of questions, each of which exclusively reflects the underlying trait or using a technical measurement term acts as an item indicator. By analyzing responses to the items, an estimate of the trait can be obtained. For this estimate to be unbiased and have low variability as a result of chance, psychometricians investigate the properties of items through the assessments of validity and reliability. In this article, we review the steps of traditional psychometric testing, as they arise from classical test theory, and more modern test procedures, based on the item response theory, as well as some statistical issues that arise in the implementation of these processes. Although the methods of classical test theory are almost always used in the development and testing of outcome instruments, methods derived from item response theory are sometimes overlooked. We briefly review the classical methods and focus on the item response theory based methods and the aspects of validity that these methods can establish.

Methods one would use when performing psychometric testing vary based on the intended use of the tool. For example, if a tool is to be used in a population that is heterogeneous with respect to patient, disease, or treatment characteristics, then samples used for psychometric testing should be representative of the target population, and differential responses to items according to patient, disease, or treatment characteristics need to be investigated. If a tool is to be used to determine the effectiveness of treatments, then the evaluation of responsiveness is crucial. Finally, if the instrument is administered through different modes such as paper and pencil, telephone, or web, the mode of administration could affect measurement properties. Although this article focuses on the validity and reliability testing, the evaluation of these measurement properties does not exhaust the list of important considerations. Developing interpretability, guidelines, and determination of minimally clinically important difference are critical to the use of the instrument in research and clinical practice and should be addressed following the reliability and validity testing.

Where Are We Now?

Psychometric Testing: Classical Test Theory

Classical test theory of instrument development considers a person's true score for the trait to be unknown. To estimate the true score, a set of items is administered to a person, and based on the item responses, the observed score is computed. This observed score is not equal to the true score because of the error of measurement [27]. For the ideal instrument, this error should be as small as possible. One situation when this error arises is when the score derived from the item responses is not exactly representative of the trait that was intended. This error is addressed by validity testing. A second source of measurement error is the chance variation in item responses, which is evaluated with reliability testing.

Although the concepts of validity and reliability in psychometric testing are grounded in classical test theory, these terms have been used more broadly. For instance, the internal and external validity are relevant to all studies and not just instrument development. Internal validity refers to the validity of inferences drawn from a study, and the use of appropriate statistical methods is a necessary (but not sufficient) condition for achieving it. External validity, on the other hand, reflects the extent to which the findings from the study can be generalized to new studies and populations. Each measurement tool is only applicable for use with a defined population. Psychometric testing is done using a sample, and this sample should be representative of the target population. Inadequate representation of the target population by the sample is one of the main sources of error that is separate from the error of measurement and from the sampling error. In the context of instrument development, internal and external validity are important for application of an instrument in assessing new target populations.

Content validity reflects the adequacy of the instrument in quantifying the underlying trait. Design validity is the adequacy of the design of the study in which the new measure is developed and tested. The sampling design should be adequate in the sense that it should produce a sample that is representative of the target population. Furthermore, an important part of the design is the timing of longitudinal assessments or timing of the collection of cross-sectional data, which could be justified by the characteristics of treatment. For example, in initial testing of the new Knee Society Scoring System, responses were collected from patients who were scheduled to undergo TKA, 6 months before the procedure, and patients who were at least 12 months post-TKA [30]. The period of 12 months was selected for postoperative followup because of the known course of improvement of pain and function experienced by patients after TKA.

Another type of validity is criterion-related validity, which exists when scores derived from an instrument reflect the underlying construct. Criterion-related validity is established in one of the four forms: (1) concurrent validity (association between two instruments measuring the same construct); (2) convergent or divergent validity (association between instrument measuring the construct of interest and a similar yet different construct); (3) dimensionality of the underlying trait (evaluated using factor analysis); and (5) predictive validity (the ability of a new measure to predict future events even when the mechanism of predictive relationship is not known).

In practice, these forms of validity are established by evaluating the magnitude of the correlations between the scores derived from different instruments. Furthermore, evidence of construct validity comes from the ability of the instrument to discriminate two or more groups of patients that are known to be different. The appropriate statistical tests such as t-tests, Fisher's exact tests, analysis of variance, chi-square test, or nonparametric tests can be used in practice, depending on the types of scores and their distributions and available sample sizes. For example, an instrument could discriminate between preoperative and postoperative groups of patients. In a longitudinal sample, the ability of the instrument to capture change over time within the same group of patients would be tested by following up patients from the preoperative to postoperative period and evaluating changes in the scores for the domains. The responsiveness of a new instrument can be compared with the existing measures and reported using summary statistics such as relative validity coefficient (ratio of two F-test statistics) [24, 29]; standardized effect size (difference between two means in SD at baseline units) [26]; and standardized response mean (SRM, difference between means in SD of the differences units) [2, 15, 26]. Larger values of these summary statistics correspond to greater responsiveness. Interpretations similar to Cohen's standardized effect sizes [8] exist (eg, SRM: $\geq$ 1.0: excellent; 0.80–0.99: good; 0.50–0.79: moderate; and $<$ 0.5: weak). As with standardized effect sizes, caution is warranted in the interpretation of these normative cutoffs [15, 41]. If an instrument shows great responsiveness but is lengthy, the tradeoff between respondent burden and responsiveness may lead to a choice of a shorter instrument in research and clinical practice.

Methods for assessment of reliability deal with the measurement error that is the result of chance and include evaluation of the internal consistency reliability (Cronbach's alpha) [11] and stability (test-retest correlations). If the true score were known, one could determine the correlation between the true score and the observed score. Because the true score is not known, a correlation coefficient cannot be computed, but the squared correlation coefficient could be bounded from below by Cronbach's alpha. Drawing a heuristic analogy with regression, where the R^2 gives the percent of variation in response explained by the explanatory variable, the observed score explains at least alpha amount of variation in true score. In practice, values of Cronbach's alpha larger than 0.7 or 0.8 are generally preferred for group level measurements, and values of 0.9 or above are preferred for measurements relating to individual cases. Thus, if clinical decisions for an individual patient are to be made based on the observed score, then higher value of Cronbach's alpha is desired [36]. Similar to the normative values for effect sizes and responsiveness coefficients, these cutoffs should be applied with care. For example, Cronbach's alpha can be inflated by the presence of a large number of items in a scale [27]. Test-retest reliability is a correlation between two observed scores obtained from the administration of the instrument at two different times. The interval between these two times should be short enough to ensure that the true value does not change and long enough for the responders not to remember their previous answers. Higher values of this correlation (eg, 0.7 or higher) indicate greater stability of scores.

Where Do We Need to Go?

Item Response Theory and Differential Item Functioning

Item response theory and differential item functioning analyses have not been frequently performed as part of instrument development and testing. One of the simplest analyses of the item response theory, Rasch analysis, has been performed for the WOMAC Osteoarthritis [50] and for the osteoarthritis of knee and hip quality of life [14]. Item response theory has also been applied to validate the use of the Subjective Knee Form developed by the International Knee Documentation Committee [19] and the International Classification of Functioning, Disability and Health [35]. This approach has also been used to develop prototypes of new measures [23] and for the evaluation of existing instruments [18, 50]. These experiences support the belief that the broader use of item response theory and differential item functioning analyses will be beneficial in the development and testing of new and existing measures in orthopaedics.

Item response theory can be applied to evaluate the extent to which a set of items is successful in measuring, indirectly, an underlying trait or construct. Using item response theory, mathematical models are developed from patients' responses to different items on a questionnaire. Different numerical values, called "parameters," are generated by the mathematical analysis and quantify the properties of each item. The different models used in item response theory are classified as one-, two-, and three-parameter, according to the number of parameters used to describe each item. The first parameter, which is present in all three models, is the item difficulty, or item location parameter. The second parameter is the item discrimination parameter. In one-parameter models, like the Rasch model, it is assumed that each item is equally discriminating, that is that all item discrimination parameters are equal [27]. For this to be true, each item must contribute equally to the separation of respondents who have a high value of the underlying trait in a particular range (around the item location) from those who have a low value. Frequently, in applications within health care, the assumption of equal discrimination is not satisfied; therefore, two-parameter models have been used [36, 38]. The third parameter, called "guessing," is not applicable in most health applications. A clear and comprehensive plan for the analysis of measurement instruments using item response theory is presented by Reeve et al. as a framework for instrument development [36].

Item bias, also known as differential item functioning (DIF), presents a major threat to the validity of measurements. Item biases can lead to inaccurate estimates of prevalence of symptoms and complications and inadequate estimates of treatment effects. Item bias occurs when responses to an item differ between two groups of respondents, for example, males and females, although there are no differences in the underlying construct (eg, severity of pain). In this case, the response to an item functioning differentially is affected by the sex of the respondent and not just the underlying construct of pain severity. Although differential item functioning may be tested using a variety of statistical approaches, the requirement for its absence can be elegantly stated in terms derived from item response theory terms: item parameters are properties of items and not properties of people who respond to items. Item response theory also offers methods for adjusting for item bias based on co-calibration of items, for example, by using different item parameters for males and females if item bias by sex is found.

Differential item functioning has been reported after analysis of responses to many widely used instruments including the European Organization for Research and Treatment of Cancer core Quality of Life Questionnaire [25], the Hospital Anxiety and Depression Scale [33], measures of physical functioning [34, 45], and the General Health Questionnaire [13]. In addition, work emerging from the Patient Reported Outcome Measurement Information System contains reports of DIF in depression items and recommendations for further testing [46].

An additional source of item bias is variation in the response of an instrument with time. This is of critical importance because symptoms and functional impairment persist over time, and as time progresses, the responses to items may change even if the underlying construct remains constant. This effect is known as adaptation or response shift and could present a major problem for valid measurement. Item bias with respect to time jeopardizes the evaluation of treatments or interventions because item bias may be mistaken for improvement as a result of a treatment or intervention.

Differential Item Functioning From Different Modes of Administration of the Assessment

Various methods are used for completion of outcome instruments, including patient self-administration, telephone interviews, automated voice response (AVR) systems, reporting through electronic devices, patient online self-reporting. Previous studies have shown that patients' responses to items may in fact vary as a function of the method of administration of the same instrument, creating "mode effects" [17, 27]. The rise in computer and Internet use makes web-based assessment an inexpensive and accessible mode of administration of instruments. The computerized modes of data collection such as the AVR system, personal digital assistants (PDAs), or web-based reporting have been compared with paper and pencil self-administration and live telephone interviews [1, 7, 12, 17, 28, 40, 43, 49]. Comparisons of a paper and pencil administration to a PDA had varying results: mode effects were found for the Center for Epidemiological Studies Depression scale [43] with higher scores in paper and pencil administration compared with in-person interviews [7] but not for measures of quality of life [1], attitudes [12], and interest [17]. In patients with cancer, higher symptom severity scores were elicited in response to automated telephone monitoring compared with live telephone interviews [40].

It is crucial that the potential effects of mode of administration be considered when data are collected using different modes of administration of an instrument. For example, in the case of a multisite study in which access to technology for the assessments varies by site, the study findings may be affected by the mode of instrument administration. Another example of mode effects comes from a situation in which different types of supportive care interventions are delivered using different modes of administration such as the Internet or telephone. Furthermore, mode effect may differ according to patient age [40]. Because TKA is performed in patients of a broad range of ages [30], mode effects may affect the determination of treatment effects if an outcome instrument is made available in a variety of forms (eg, Internet-based and pencil and paper-based) for collection of responses.

How Do We Get There?

Testing for Differential Item Functioning

Several approaches can be used to detect differential item functioning during the development and testing of new outcome instruments. Differential item functioning is detected by comparing the item responses of people who have the same underlying true value of the construct. Because this true value is not known, it is estimated using different methods, for example, the summed score across all items, or an estimate based on item response theory [31, 32, 34, 44–46]. The analyses can be carried out in two stages: initially with all items and then without items exhibiting bias [10, 16]. One method for testing for DIF is based on the logistic regression method using a model in which each item response is related to membership of a group reflecting the bias factor in question, the summed item score, and their interaction. Nonuniform DIF (ie, item bias that varies with the value of the construct) is present if the interaction term is significant. Uniform DIF is present when the additive group effect is significant. The item response theory approach contrasts a model that assumes equality of item parameters between groups to an augmented model constructed by removing equality constraints across groups. The likelihood ratio test is used to assess the statistical significance of potential DIF detected by this method.

An inherent issue in the assessment of DIF is the effect of multiple testing, because each item in an instrument is tested at least twice. If the null hypothesis is correct, and the threshold for rejecting it is set at 5%, then five of 100 tests are expected to result in incorrect rejection of the null hypothesis due to chance. Thus, when a large number of tests is performed, there is a danger that some comparisons will be classified as significant as a result of chance. Statistical methods to deal with the multiple testing issues are available. The first method is the application of Bonferroni-type procedures to control the family-wise Type I error rate (eg, the probability of rejecting at least one correct null hypothesis in a family of multiple hypotheses). Although the Bonferroni approach does not require any additional assumptions, it is often too stringent leading to inflation of the Type II error. The second method controls for the false discovery rate using Benajmini-Hochberg or Hochberg adjustments under the appropriate assumptions [4, 5, 20, 47].

Adjustment for Differential Item Functioning

When item bias is present, adjustment for its influence on the outcome score cannot be made simply by including the relevant variable (such as a patient characteristic) as a covariate in a statistical model. This approach will combine differences resulting from item bias with the actual differences present between groups of patients and therefore will not flag DIF. Another approach, based on item response theory, is to adjust the item parameters on the basis of membership of the groups for which differential item response has been detected. In deciding on the need of the adjustment, the evaluation of the magnitude of DIF is critical, because item bias may exist, as indicated by statistical significance tests, but may not be clinically or practically important. Although much work has been done in finding consensus about clinical significance [15, 41, 42], the discussion of the clinical and practical significance of DIF has been lacking in the literature. At the stage of instrument development, biased items may be removed, especially if additional subject matter considerations support removal of item in question.

In summary, methods of validity analysis based on item response theory and DIF allow us to extend the evaluation of outcome instruments beyond the limitations of classical test theory. The application of these methods could improve the quality of instruments used in orthopaedics, increasing the validity of outcome measures. This advance is expected to contribute to improvements in decision-making and maintenance of effective patient care.

References

1. Agel J, Rockwood T, Mundt JC, Greist JH, Swiontkowski M. Comparison of interactive voice response and written self-administered patient surveys for clinical research. *Orthopedics.* 2001;24:1155–1157.
2. Beaton D, Hogg-Jonson 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;50:79–83.
3. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt L. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes following total hip or knee arthroplasty in osteoarthritis. *J Orthopaedic Rheumatol.* 1988;1:95–108.
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B.* 1995;57:289–300.
5. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat.* 2001;29:1165–1188.
6. Bourne RB, Chesworth BM, Davis AM, Mahomed NN, Charron KD. Patient satisfaction after total knee arthroplasty: who is satisfied and who is not? *Clin Orthop Relat Res.* 2010;468:57–63.
7. Chan KS, Orlando M, Ghosh-Dastidar B, Duan N, Sherbourne CD. The interview mode effect on the Center for Epidemiologic Studies Depression (CES-D) scale: an item response theory analysis. *Med Care.* 2004;42:281–289.
8. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ, USA: Erlbaum; 1988.

9. Collins NJ, Roos EM. Patient-reported outcomes for total hip and knee arthroplasty: commonly used instruments and attributes of a 'good' measure. *Clin Geriatr Med.* 2012;28:367–394.
10. Crane P, Gibbons L, Narasimhalu K, Lai JS, Cella D. Rapid detection of differential item functioning in assessments of health-related quality of life: the functional assessment of cancer therapy. *Qual Life Res.* 2007;16:101–114.
11. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrica.* 1951;16:297–334.
12. Donovan MA, Drasgow F, Probst TM. Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *J Appl Psychol.* 2000;85:305–313.
13. Gao W, Stark D, Bennet MI, Siegert RJ, Murray S, Higginson IJ. Using the 12-item General Health Questionnaire to screen psychological distress from survivorshp to end-of ife care: dimensionality and item quality. *Psychooncology.* 2012;21:954–961.
14. Goetz C, Ecosse E, Rat AC, Pouchot J, Coste J, Guillemin F. Measurement properties of the osteoarthritis of knee and hip quality of life OAKHQOL questionnaire: an item response theory analysis. *Rheumatology (Oxford).* 2011;50:500–505.
15. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis.* 1987;40:171–178.
16. Hambleton R. Good practices for identifying differential item functioning. *Med Care.* 2006;44(Suppl 3):S182–S188.
17. Hansen JI, Neuman JL, Haverkamp BE, Lubinski BR. Comparison of user reaction to two methods of strong interest inventory administration and report feedback. *Measurement and Evaluation in Counseling and Development.* 1997;30:115–127.
18. Hart DL, Deutscher D, Crane PK, Wang YC. Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. *Qual Life Res.* 2009;18:1067–1083.
19. Higgins LD, Taylor MK, Park D, Ghodadra N, Marchant M, Pietrobon R, Cook C; International Knee Documentation Committee. Reliability and validity of the International Knee Documentation Committee (IKDC) Subjective Knee Form. *Joint Bone Spine.* 2007;74:594–599.
20. Holland PW, Rosenbaum R. Conditional association and unidimensionality in monotone latent variable models. *Ann Stat.* 1986;14:1523–1543.
21. Impellizzeri FM, Mannion AF, Leunig M, Bizzini M, Naal FD. Comparison of the reliability, responsiveness, and construct validity of 4 different questionnaires for evaluating outcomes after total knee arthroplasty. *J Arthroplasty.* 2011;26:861–869.
22. Insall JN, Dorr LD, Scott RD, Scott WN. Rationale of the Knee Society Clinical Scoring System. *Clin Orthop Relat Res.* 1989;248:13–14.
23. Jette AM, McDonough CM, Haley SM, Ni P, Olarsch S, Latham N, Hambleton RK, Felson D, Kim YJ, Hunter D. A computer-adaptive disability instrument for lower extremity osteoarthritis research demonstrated promising breadth, precision, and reliability. *J Clin Epidemiol.* 2009;62:807–815.
24. Kosinski M, Bjorner J, Ware J, Batenhorst A, Cady R. The responsiveness of headache impact scales scored using 'classical' and 'modern' psychometric methods: a re-analysis of three clinical trials. *Qual Life Res.* 2003;12:903–912.
25. Lheureux M, Raherison C, Vernejoux JM, Nguyen L, Nocent C, Tunon De Lara M, Taytard A. Quality of life in lunch cancer: does disclosure of the diagnosis have an impact? *Lung Cancer.* 2004;43:175–182.
26. Liang M, Fossel A, Larson M. Comparison of five health status instruments for orthopedic evaluation. *Med Care.* 1990;28:632–642.
27. Lord F. *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates; 1980.
28. Mead AD, Drasgow F. Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychol Bull.* 1993;114:449–458.
29. McHorney C, Haley S, Ware J. 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;50:451–461.
30. Noble PC, Scuderi GR, Brekke A, Sikorskii A, Benjamin J, Lonner J, Chadha P, Daylamani D, Scott WN, Bourne RB. Development of a New Knee Society Scoring System. *Clin Orthop Relat Res.* 2012;470:20–32.
31. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. *Appl Psychol Meas.* 2002;24:50–64.
32. Orlando M, Thissen D. Further examination of the performance of S-Chi-Square, an item fit index for dichotomus item response theory models. *Appl Psychol Meas.* 2003;27:289–298.
33. Osborne RH, Elsworth GR, Sprangers MA, Oort FJ, Hopper JL. The value of the Hospital Anxiety and Depression Scale (HADS) for comparing women with early onset breast cancer with population-based reference women. *Qual Life Res.* 2004;13:191–206.
34. Petersen M, Groenvold M, Bjorner J, Aaronson N, Conroy T, Cull A, Fayers P, Hjermstad M, Spraingers M, Sullivan M; European Organisation for Research and Treatment of Cancer Quality of Life Group. Use of differential item functioning to assess the equivalence of translations of a questionnaire. *Qual Life Res.* 2003;12:373–385.
35. Pollard B, Dixon D, Dieppe P, Johnston M. Measuring the ICF components of impairment, activity limitation and participation restriction: an item analysis using classical test theory and item response theory. *Health Qual Life Outcomes*. 2009;7:41.
36. Reeve BB, Hays RD, Bjorner JL, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai JS, Cella D; PROMIS Cooperative Group. Psychometric evaluation and calibration of health-related quality of life item banks. *Med Care.* 2007;45(Suppl 1):S22–S26.
37. 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.
38. Samejima F. Graded response model. In: van der Linden WJ, Hambleton RK, eds. *Handbook of Modern Item Response Theory*. Berlin, Germany: Springer; 1997:85–100.
39. Scuderi GR, Bourne RB, Noble PC, Benjamin JB, Lonner JH, Scott WN. The New Knee Society Scoring System. *Clin Orthop Relat Res*. 2012;470:3–19.
40. Sikorskii A, Given C, Given B, Jeon S, You M. Differential symptom reporting by mode of administration of the assessment: automated voice response system versus a live telephone interview. *Med Care.* 2009;47:866–874.
41. Sloan JA, Berk L, Roscoe J, Fisch MJ, Shaw EG, Wyatt G, Morrow GR, Dueck AC; National Cancer Institute. Integrating patient-reported outcomes into cancer symptom management clinical trials supported by the National Cancer Institute-sponsored clinical trials networks. *J Clin Oncol.* 2007;25:5070–5077.
42. Sloan JA, Cella D, Hays RD. Clinical significance of patient-reported questionnaire data: another step toward consensus. *J Clin Epidemiol.* 2005;58:1217–1219.
43. Swartz RJ, de Moor K, Cook KF, Fouladi RT, Basen-Engquist K, Eng C, Carmack Taylor CL. Mode effects in the Center for Epidemiologic Studies Depression (CES-D) scale: personal digital assistant vs. paper and pencil administration. *Qual Life Res.* 2007;16:803–813.
44. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages, and some neglected topics. *Med Care.* 2006;44(Suppl 3):S152–S170.
45. Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, Morales LS, Orlando-Edelen M, Cella D.

Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Qual Life Res.* 2007;16(Suppl 1):43–68.

46. Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, Lai JS, Choi SW, Hays RD, Reeve BB, Reise SP, Pilkonis PA, Cella D. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): an item response theory approach. *Psychol Sci Q.* 2009;51:148–180.
47. Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamimi-Hochberg procedure for controlling the false positive reate in multiple comparisions. *J Educ Behav Stat.* 2002;27:77–83.
48. Wang D, Jones MH, Khair MM, Miniaci A. Patient-reported outcome measures for the knee. *J Knee Surg.* 2010;23:137–151.
49. Weiler K, Christ AM, Woodworth CG, Weiler RL, Weiler JM. Quality of patient-reported outcome data captured using paper and interactive voice response diaries in an allergic rhinitis study: is electronic data capture really better? *Ann Allergy Asthma Immunol.* 2004;92:335–339.
50. Wolfe F, Kong SX. Rasch analysis of the Western Ontario MacMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. *Ann Rheum Dis.* 1999;58:563–568.