한국전문물리치료학회지 2012년 19권 3호
Phys Ther Kor. Vol. 19 No. 3 2012

Computer Adaptive Testing Method for Measuring Disability in Patients With Back Pain

Bongsam Choi[1], PhD, MPH, PT
[1]Dept. of Physical Therapy, College of Health and Welfare, Woosong University

Abstract

Most conventional instruments measuring disability rely on total score by simply adding individual item responses, which is dependent on the items chosen to represent the underlying construct (test-dependent) and a test statistic, such as coefficient alpha for the estimate of reliability, varying from sample to sample (sample-dependent). By contrast, item response theory (IRT) method focuses on the psychometric properties of the test items instead of the instrument as a whole. By estimating probability that a respondent will select a particular rating for an item, item difficulty and person ability (or disability) can be placed on same linear continuum. These estimates are invariant regardless of the item used (test-free measurement) and the ability of sample applied (sample-free measurement). These advantages of IRT allow the creation of invariantly calibrated large item banks that precisely discriminate the disability levels of individuals. Computer adaptive testing (CAT) method often requiring a testing algorithm promise a means for administering items in a way that is both efficient and precise. This method permits selectively administering items that are closely matched to the ability level of individuals (measurement precision) and measuring the ability without the loss of precision provided by the full item bank (measurement efficiency). These measurement properties can reasonably be achieved using IRT and CAT method. This article aims to investigate comprehensive overview of the existing disability instrument for back pain and to inform physical therapists of an alternative innovative way overcoming the shortcomings of conventional disability instruments. An understanding of IRT and CAT method will equip physical therapist with skills in interpreting the measurement properties of disability instruments developed using the methods.

[Bongsam Choi. Computer Adaptive Testing Method for Measuring Disability in Patients With Back Pain. Phys Ther Kor. 2012;19(3):124-131.]

Key Words: Computer adaptive testing; Disability; Item response theory; Low back pain.

Introduction

Measuring disability is crucial in capturing clinical changes, evidence-based rehabilitation practice, administration of disability management, and policy making process. Due to the fact that back pain is the most common cause of activity limitation in our society, measuring disability resulting from back pain has become an important issue (Andersson, 1999). This has prompted extensive research on self-report disability instruments for back pain (Bergner et al, 1981; Daltroy et al, 1996; Davidson and Keating, 2002; Fisher, 1999; Fritz and Irrgang, 2001; Kopec et al, 1996; McHorney et al, 1997; Roland and Morris, 1983a; Ware et al, 1996; Ware and Sherbourne, 1992; Williams and Myers, 1998a). To date, nearly 82 back-related disability instruments for back pain have been identified in review of the major medical databases (Müller et al, 2006). Most, if not all, of these instruments in peer reviewed journals appeared to have adequate psychometric properties based on various criteria such as validity, internal consistency, responsiveness to changes and floor/ceiling effects (Kopec, 2000). However, only a few instruments have

Corresponding author: Bongsam Choi bchoi@wsu.ac.kr

been widely used and commonly accepted for measuring disability resulted from back pain (Kopec et al, 1996; Kopec, 2000; Müller et al, 2006).

Despite the myriad of disability instruments for back pain available, selecting an optimal instrument is a prevailing challenge. The challenge stems from the need to carefully consider the preferences of investigators or clinician when choosing an instrument (Kopec, 2000). Although those previously developed instruments have adequate psychometric properties, they may or may not be sensitive to all severity groups or to the actual improvements that result from particular rehabilitation interventions. These instruments are often developed to target the "average" person, therefore they tend to be more sensitive at the center than at the extremes (e.g., low and high levels of disability) of the ability (or disability) range (McHorney, 1997). For example, while the Oswestry Back Pain Disability Questionnaire (ODQ), is often considered as a gold standard, it often demonstrate ceiling effects (i.e., persons with high ability) when it is administered to persons with minimal impairments (Fairbank and Pynsent, 2000; White and Velozo, 2002) and floor effects (i.e., person with low ability) when it is administered to persons with severe impairments (Deyo et al, 1998). Thus the ODQ often fails to precisely measure the disability of back pain across the full range of ability.

The purpose of this study is to review existing disability instruments for back pain, disadvantages of using the instruments, and introduce an innovative way. That is, computer adaptive testing (CAT) combined with item response theory (IRT) methodology in measuring disability for back pain.

Methods

Self-reported outcome measures

Self-reported outcome measures are generally classified as generic or condition-specific measure (McHorney, 1997; McHorney, 1999). Generic measures often include global ratings of health status as well as ratings of multi dimensional status of health-related quality of life. These instruments often measure a broad spectrum of health concepts and are intended to provide scores that are sensitive to disease severity. By contrast, condition-specific measures are designed to assess the aspects of health status affected by certain disease pathology and view the attribution of symptom and functional limitations to a specific condition (Kopec, 2000). Thus, in contrast to generic measures, condition-specific measures are likely to be sensitive to treatment and natural history of a specific disease or condition.

Although generic measures were not primarily designed to assess the specific conditions, two instruments, the Sickness Impact Profile (SIP) and the Physical Function Scale (PF-10) have been applied to chronic back pain. The SIP was originally developed and validated as a measure of sickness-related behavioral dysfunction consisting of 189 items in 14 categories (Carter et al, 1976). With few revisions, the final version of the SIP was developed as a behavioral-based measure of health status for use in a variety of chronic diseases (Deyo, 1986). The PF-10 is a subscale of the short form (SF)-36 that measures physical functioning, which assesses limitations in a variety of physical activities. Other versions of PF-10, such as a general population version PF-12 PCS (Physical Component Summary) and specific low back version Physical Functioning (PF)-18 (Davidson and Keating, 2002) have been developed. Among patients with back pain, studies report adequate psychometric properties for these two instruments (Bergner et al, 1981; Ware and Sherbourne, 1992; Ware et al, 1996).

As disease specific measures for back pain, the Roland-Morris Disability Questionnaire (RMDQ), the Quebec Back Pain Disability Scale (QBDS), and the ODQ are the most widely accepted instruments. The RMDQ consists of 24 items of daily physical activity from the SIP. In contrast to the SIP, the RMDQ is short, simple to complete, and readily understood by

patients (Roland and Morris, 1983a). The QBDS consists of 20 items of a comprehensive view of person's disability for back pain, which adopted the World Health Organization's International Classification of Functioning, Disability and Health (ICF) as a conceptual model to select test items relevant to ICF activity and participation domains (Kopec et al, 1995; Kopec et al, 1996). One of unique features about the QBDS is that it measures only physical function domain, while most instruments appear to assess more than one domain within the assessment (Kopec et al, 1995). All of them appear to have good psychometric properties supported by many studies (Beurskens et al, 1996; Kopec and Esdaile, 1995; Müller et al, 2006; Roland and Morris, 1983a; Stratford et al, 1996a).

The Oswestry Back Pain Disability Questionnaire as a gold standard

The ODQ was first introduced by John O'Brien in 1976 and further developed by Fairbank and colleagues in 1980 (Fairbank and Pynsent, 2000). The ODQ consisted of 10 items assessing the level of pain and interference with personal care, physical activities (i.e., lifting, walking, sitting, and standing, sleeping, sex life, social life, and traveling. Its several validated versions have also been published omitting a single item (i.e., sex life or social life) (Bossons et al, 1996) or replacing "sex life" item with employment/homemaking item (Fritz and Irrgang, 2001). The ODQ and its revised versions have been proved to be much more sensitive to patients with severe symptoms, while they also appear to be occasionally responsive to those with minor symptoms (Fairbank and Pynsent, 2000). The ODQ, whether in the original or revised versions, remains a salient measure of condition-specific disability with good validity and reliability (Fairbank et al, 1980; Fairbank, 2000a; Fairbank and Pynsent, 2000; Fritz and Irrgang, 2001; Müller et al, 2006; White and Velozo, 2002). The ODQ is one of the most widely accepted back pain-specific instruments (Fairbank, 2007; Frost et al, 2008; Kopec, 2000; White and Velozo, 2002). It is presently considered as the "gold standard" in the assessments of back pain (Fairbank and Pynsent, 2000) because of its many advantages such as popularity, internally consistent scale, good reliability and responsiveness to clinical change. In numerous studies, the ODQ and the revised versions of it are recommended as a standardized measure of physical function in individuals with back pain (Davidson and Keating, 2002; Fairbank and Pynsent, 2000; Fritz and Irrgang, 2001; Frost et al, 2008; Kopec, 2000; Kopec and Esdaile, 1995; Müller et al, 2006; Stewart, 2003; Taylor et al, 1999; White and Velozo, 2002).

Despite the popularity of ODQ in health care, there have been a few concerns about several of its measurement properties. The ODQ is shown to be the multidimensional construct. Physical function and pain item as separate construct (Page et al, 2002; White and Velozo, 2002) and lacks of sensitivity to reliably discriminate individuals in particular ranges of the scale due to "gaps" between test items (e.g., none of items were available a gap between "standing" and "lifting" on item difficulty hierarchical order) for the underlying continuum (White and Velozo, 2002). The lack of breadth may lead to inadequate sensitivity at the extremes of the scale. Not surprisingly, the developers of ODQ and researchers indicate that the instrument is better at detecting change only in a specific disability level due to its substantial measurement imprecision (Fairbank, 2007; Fairbank and Pynsent, 2000). Despite these limitations, the ODQ remains a leading back pain disability instrument in health care (Davidson and Keating, 2002; Deyo et al, 1998; Fairbank and Pynsent, 2000; Fairbank, 2007; Fritz and Irrgang, 2001; Kopec and Esdaile, 1995; Müller et al, 2006; Stewart, 2003; Taylor et al, 1999; White and Velozo, 2002).

Shortcomings of existing disability measures

Measurement imprecision such as ceiling effects result in type II errors. That is, the number of false-negatives is large among those scoring in the

upper extreme of the instrument (McHorney, 1997). Furthermore, it is impossible to measure improvement in health status over time for those in the ceiling (i.e., those able to complete all items without difficulty or scoring high initially). These limitations in measurement precision are partly due to the fixed number of items included on instruments. Thus, these instruments do not have adequate breadth for the underlying construct being measured (Liang et al, 2002). This subsequently leads to ceiling/floor effects and failure to capture the small improvements while they are potentially significant increments of improvement across the full ranges of the construct (McHorney et al, 1997; Velozo et al, 1999).

Measurement imprecision may also be the result of using items that do not closely match to the ability of the population of interest (McHorney, 1999). Deficits in precision occur when easy items are administered to high ability populations (e.g., administering items measuring basic activities of daily living to elite athletes) and difficult items are administered to low ability populations (e.g., administering items measuring ability to lift heavy objects to individuals with severe back pain). Furthermore, since individuals are asked to respond all items on an instrument regardless of how they respond to previous items (i.e., asked if they can lift 10 pounds after responding that they cannot lift 5 pounds) and regardless of the relevance of the items to the individual (i.e., asking individual with no movement in legs if they can walk a mile), respondent burden and administration costs are increased. Unfortunately, test-level statistics provide little insight in regards to how to eliminate these limitations. These limitations are a function of characteristics of the classical test theory (CTT) model.

Despite the popularity and widespread use of instruments developed using the CTT model, existing disability measures have numerous shortcomings (Jette and Haley, 2005). In general, disability instruments created under the CTT paradigm yield total scores obtained by adding individual item responses. These scores provide only a general sense of a person's ability level (i.e., disability level) and often fail to provide detailed item level psychometrics (i.e., no detail information is provided about how an individual performs on each item). First, the total score is dependent on the items chosen to represent the underlying construct (test-dependent) (Hambleton, 2000). That is, respondents will have lower scores on difficult items and higher scores on easier items while their ability remains the same. Second, the test scores obtained from a sample cannot be compared across different samples (sample-dependent) (DeVellis, 2006; Hambleton, 2000). That is, test statistics such as coefficient alpha for the estimate of reliability or correlations for estimates of vary from sample to sample (i.e., sample dependent). Third, test scores are non-linear summed scores, which yield ordinal raw scores (Wright and Linacre, 1989). These ordinal scores may be insensitive to changes at the extremes of the scale (Wright and Linacre, 1989).

Item response theory and computer adaptive testing

In contrast to the CTT, the IRT focuses on the psychometric properties of the items making up instrument instead of the instrument as a whole (Velozo et al, 2006). By estimating the probability that a respondent will select a particular rating for an item, item difficulty and person ability (or disability) can be placed on the same linear continuum. Thus, IRT model allows "connecting" individuals' responses to items at their ability level (Velozo et al, 2006; Velozo and Peterson, 2001). Estimates of person ability (or disability) on an underlying construct obtained using IRT methods are invariant regardless of the items used (i.e., test free measurement), whereas under the CTT paradigm, person scores vary depending on the difficulty of the instrument. Furthermore, item difficulty estimates derived from the IRT analyses are independent of the ability of the sample (i.e., sample free measurement), while test statistics in CTT are dependent on the sample tak-

ing the test. In addition, the Rasch model (one-parameter IRT model) can linearly transform raw scores (typically used in analyses based on CTT) into equal interval measures (Velozo et al, 1999). These advantages of IRT allow for the creation of invariantly calibrated large item banks that can more precisely discriminate individuals' ability levels and thus, capture smaller increments of change.

In order to achieve the goal of measurement precision, a disability measure should have items covering the full range of the underlying construct and capturing the small increments of changes (Jette and Haley, 2005). With optimal measurement precision, one can theoretically yield measures of equal precision at all levels of the underlying construct, thus achieving what has been termed equiprecise measurement (Weiss, 1982). That is, the measure is capable of measuring a wide range of disability from the least able (or most disabled) to the most able (or least disabled). Unlike the existing "fixed" disability assessments that require all the items of an instrument, equiprecise measurement fosters item selection determined by disability level. For example, when measuring the physical function of a person with mild back pain, items would be chosen which closely match the ability of this individuals (i.e., more difficult items would be chosen). Similarly, when measuring the physical function of a person with severe back pain, items will be chosen that closely match the severely impaired person. These two persons will be measured on the same physical-function scale with different sets of items (Velozo et al, 1999).

While IRT methodologies provide the means for generating and linking person ability and item difficulty calibrations, CAT methods promise a means for administrating items in a way that is both efficient and precise (Bjorner and Ware, 1998; Elhan et al, 2008; Haley et al, 2004b; Jette and Haley, 2005; McHorney, 1997; Velozo et al, 1999; Velozo et al, 2000). Studies have shown that CAT improves test efficiency maintaining adequate precisions with fewer items than the full test. Six to 7 items have been shown on average to achieve a standard error of ability estimates of .3 (Haley et al, 2008; Haley et al, 2006a; Hart et al, 2006; Jette et al, 2008).

The CAT often requires a testing algorithm which defines iterative processes with a set of rules specifying the test questions to be administered to respondents. This includes procedures for item selection, ability estimation, and termination criteria. By selectively administering items that are matched to the ability level of the individuals, measurement efficiency can be accomplished without the loss of precision provided by the full item bank. For example, when measuring the disability of a person with mild back pain, items would be chosen that matched the mildly impaired ability. Similarly, when measuring the disability of a person with more severe back pain, a different set of items would be chosen that match that individual's severely impaired ability. With this technology, a small number of items can be selected from the item bank which are most relevant and targeted to a person of a particular ability (Velozo et al, 1999).

Numerous studies have recently found that CAT improves both in measurement precision and efficiency relative to the full test (Elhan et al, 2008; Flynn et al, 2008; Haley et al, 2008; Hart et al, 2006; Hol et al, 2007; Jette et al, 2008; Ware et al, 2003; Weiss, 1982). Several studies have also been reported that CAT measures are highly correlated with other instruments measuring same construct and require fewer number of items with an average 6 items to reach the ability estimation (Fliege et al, 2009; Hart et al, 2008a; Hart et al, 2008b; Shone et al, 1993). IRT in combination with CAT has become an alternative to conventional fixed-format disability measurement (Jette and Haley, 2005; Kopec, 2000).

Conclusion

Although the psychometric property in CTT paradigm such as reliability, validity, or responsiveness

are well-known and rigorous criteria to select a proper outcome measure, the properties may not be sufficient in terms of measurement precision and efficiency. The CAT method with IRT-based measures will promise measurement efficiency and precision in measuring patient with back pain. These two methodologies can be recommended for use in measuring disability and should improve measurement quality in terms of precision and efficiency of outcome assessment for back pain.

References

Andersson GB. Epidemiological features of chronic low-back pain. Lancet. 1999;354(9178):581-585.

Bergner M, Bobbitt RA, Carter WB, et al. The Sickness Impact Profile: Development and final revision of a health status measure. Med Care. 1981;19(8):787-805.

Beurskens AJ, de Vet HC, Köke AJ. Responsiveness of functional status in low back pain: A comparison of different instruments. Pain. 1996;65(1):71-76.

Bjorner J, Ware Jr JE. Using modern psychometric methods to measure health outcomes. Med Outcome Trust Monitor. 1998;3:12-16.

Bossons CR, Levy J, Sutterlin CE 3rd. Reconstructive spinal surgery: Assessment of outcome. South Med J. 1996;89(11):1045-1052.

Carter WB, Bobbitt RA, Bergner M, et al. Validation of an interval scaling: The sickness impact profile. Health Serv Res. 1976;11(4):516-528.

Daltroy LH., Cats-Baril WL, Katz JN, et al. The North American spine society lumbar spine outcome assessment instrument: Reliability and validity tests. Spine (Phila Pa 1976). 1996;21(6):741-749.

Davidson M, Keating JL. A comparison of five low back disability questionnaires: Reliability and responsiveness. Phys Ther. 2002;82(1):8-24.

DeVellis RF. Classical test theory. Med Care. 2006;44(11 Suppl 3):S50-S59.

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(9):951-954.

Deyo RA. Measuring the functional status of patients with low back pain. Arch Phys Med Rehabil. 1988;69(12): 1044-1053.

Deyo RA, Battie M, Beurskens AJ, et al. Outcome measures for low back pain research. A proposal for standardized use. Spine (Phila Pa 1976). 1998;23(18):2003-2013.

Elhan AH, Oztuna D, Kutlay S, et al. An initial application of computerized adaptive testing (CAT) for measuring disability in patients with low back pain. BMC Musculoskelet Disord. 2008;9:166.

Fairbank JC. Use and abuse of Oswestry Disability Index. Spine (Phila Pa 1976). 2007;32(25):2787-2789.

Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976). 2000;25(22):2940-2952.

Fisher WP Jr. Foundations for health status metrology: The stability of MOS SF-36 PF-10 calibrations across samples. J La State Med Soc. 1999;151(11):566-578.

Fliege H, Becker J, Walter OB, et al. Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. Int J Methods Psychiatr Res. 2009;18(1):23-36.

Flynn KE, Dombeck CB, DeWitt EM, et al. Using item banks to construct measures of patient reported outcomes in clinical trials: Investigator perceptions. Clin Trials. 2008;5(6):575-586.

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(2):776-788.

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(22):2450-2457.

Haley SM, Coster WJ, Andres PL, et al. Score comparability of short forms and computerized adaptive testing: Simulation study with the activity measure for post-acute care. Arch Phys Med Rehabil. 2004b;85(4):661-666.

Haley SM, Gandek B, Siebens H, et al. Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation: II. Participation outcomes. Arch Phys Med Rehabil. 2008;89(2):275-283.

Haley SM, Ni P, Ludlow LH, et al. Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. Arch Phys Med Rehabil. 2006a;87(9):1223-1229.

Hambleton RK. Emergence of item response modeling in instrument development and data analysis. Med Care. 2000;38(9 Suppl):II60-II65.

Hart DL, Cook KF, Mioduski JE, et al. Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. J Clin Epidemiol. 2006;59(3):290-298.

Hart DL, Wang YC, Stratford PW, et al. A computerized adaptive test for patients with hip impairments produced valid and responsive measures of function. Arch Phys Med Rehabil. 2008a;89(11):2129-2139.

Hart DL, Wang YC, Stratford PW, et al. Computerized adaptive test for patients with knee impairments produced valid and responsive measures of function. J Clin Epidemiol. 2008b;61(11):1113-1124.

Hol AM, Vorst HCM, Mellenbergh GJ. Computerized adaptive testing for polytomous motivation items: Administration mode effects and a comparison with short forms. Applied Psychological Measure. 2007;31:412-429.

Jette AM, Haley SM. Contemporary measurement techniques for rehabilitation outcomes assessment. J Rehabil Med. 2005;37(6):339-345.

Jette AM, Haley SM, Ni P, et al. Creating a computer adaptive test version of the late-life function and disability instrument. J Gerontol A Biol Sci Med Sci. 2008;63(11):1246-1256.

Kopec JA. Measuring functional outcomes in persons with back pain: A review of back-specific questionnaires. Spine (Phila Pa 1976). 2000;25(24):3110-3114.

Kopec JA, Esdaile JM. Functional disability scales for back pain. Spine (Phila Pa 1976). 1995;20(17):1943-1949.

Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec Back Pain Disability Scale. Measurement properties. Spine (Phila Pa 1976). 1995;20(3):341-352.

Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec Back Pain Disability Scale: Conceptualization and development. J Clin Epidemiol. 1996;49(2):151-161.

Liang MH, Lew RA, Stucki G, et al. Measuring clinically important changes with patient-oriented questionnaires. Med Care. 2002;40(4 Suppl):II45-II51.

McHorney CA. Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med. 1997;127(8 Pt 2):743-750.

McHorney CA. Health status assessment methods for adults: Past accomplishments and future challenges. Annu Rev Public Health. 1999;20:309-335.

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(4):451-461.

Müller U, Roder C, Greenough CG. Back related outcome assessment instruments. Eur Spine J. 2006;15(Suppl 1):S25-S31.

Page SJ, Shawaryn MA, Cernich AN, et al. Scaling of the revised Oswestry low back pain

questionnaire. Arch Phys Med Rehabil. 2002;83(11):1579-1584.

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). 1983a;8(2):141-144.

Shone CC, Quinn CP, Wait R, et al. Proteolytic cleavage of synthetic fragments of vesicle-associated membrane protein, isoform-2 by botulinum type B neurotoxin. Eur J Biochem. 1993;217(3):965-971.

Stewart AL. Conceptual challenges in linking physical activity and disability research. Am J Prev Med. 2003;25(3 Suppl 2):137-140.

Stratford PW, Binkley JM, Riddle DL. Health status measures: Strategies and analytic methods for assessing change scores. Phys Ther, 1996a;76(10):1109-1123.

Taylor SJ, Taylor AE, Foy MA, et al. Responsiveness of common outcome measures for patients with low back pain. Spine (Phila Pa 1976). 1999;24(17):1805-1812.

Velozo CA, Choi B, Zylstra SE, et al. Measurement qualities of a self-report and therapist-scored functional capacity instrument based on the Dictionary of Occupational Titles. J Occup Rehabil. 2006;16(1):109-122.

Velozo CA, Kielhofner G, Lai JS. The use of Rasch analysis to produce scale-free measurement of functional ability. Am J Occup Ther. 1999;53(1):83-90.

Velozo CA, Lai JS, Mallinson T, et al. Maintaining instrument quality while reducing items: Application of Rasch analysis to a self-report of visual function. J Outcome Meas. 2000;4(3):667-680.

Ware JE 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(3):220-233.

Ware JE Jr, Kosinski M, Bjorner JB, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Qual Life Res. 2003;12(8):935-952.

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(6):473-483.

Weiss D. Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement. 1982;6:473-492.

White LJ, Velozo CA. The use of Rasch measurement to improve the Oswestry classification scheme. Arch Phys Med Rehabil. 2002;83(6):822-831.

Williams RM, Myers AM. Functional Abilities Confidence Scale: A clinical measure for injured workers with acute low back pain. Phys Ther. 1998a;78(6):624-634.

Wright BD, Linacre JM. Observations are always ordinal; Measurements, however, must be interval. Arch Phys Med Rehabil. 1989;70(12):857-860.

This article was received July 16, 2012, was reviewed July 17, 2012, and was accepted August 6, 2012.