Item Response Theory, Computerized Adaptive Testing, and PROMIS: Assessment of Physical Function

James F. Fries, James Witter, Matthias Rose, David Cella, Dinesh Khanna, and Esi Morgan-DeWitt

ABSTRACT. ***Objective.*** Patient-reported outcome (PRO) questionnaires record health information directly from research participants because observers may not accurately represent the patient perspective. Patient-reported Outcomes Measurement Information System (PROMIS) is a US National Institutes of Health cooperative group charged with bringing PRO to a new level of precision and standardization across diseases by item development and use of item response theory (IRT).

Methods. With IRT methods, improved items are calibrated on an underlying concept to form an item bank for a "domain" such as physical function (PF). The most informative items can be combined to construct efficient "instruments" such as 10-item or 20-item PF static forms. Each item is calibrated on the basis of the probability that a given person will respond at a given level, and the ability of the item to discriminate people from one another. Tailored forms may cover any desired level of the domain being measured. Computerized adaptive testing (CAT) selects the best items to sharpen the estimate of a person's functional ability, based on prior responses to earlier questions. PROMIS item banks have been improved with experience from several thousand items, and are calibrated on over 21,000 respondents.

Results. In areas tested to date, PROMIS PF instruments are superior or equal to Health Assessment Questionnaire and Medical Outcome Study Short Form-36 Survey legacy instruments in clarity, translatability, patient importance, reliability, and sensitivity to change.

Conclusion. Precise measures, such as PROMIS, efficiently incorporate patient self-report of health into research, potentially reducing research cost by lowering sample size requirements. The advent of routine IRT applications has the potential to transform PRO measurement. (First Release Nov 15 2013; J Rheumatol 2014;41:153–8; doi:10.3899/jrheum.130813)

Key Indexing Terms:
ITEM RESPONSE THEORY COMPUTERIZED ADAPTIVE TESTING PROMIS
PHYSICAL FUNCTION DISABILITY

Patient-reported outcomes (PRO) measures such as the Health Assessment Questionnaire (HAQ)[1] and the 10-item physical function instrument (PF-10) derived from the Medical Outcome Study Short Form-36 Survey (SF-36)[2] have become central to evaluation of treatment and study of the disease course in rheumatic diseases over the past 30 years. Recent developments in item response theory (IRT) and computerized adaptive testing (CAT) now permit marked improvement in the effectiveness of PRO outcome assessment, often with better items used in better ways. The precision of estimation of a latent trait (such as physical function) can be improved, and the range of disease severity that can be accurately assessed can be increased, resulting in smaller sample size requirements and/or shorter questionnaires. The physical function (PF) domain has been broadened from the earlier disability domain so that function both above and below the population mean are included[1,2,3,4,5]. Change in traditional outcome assessment methodology is long overdue, and outcome assessment investigators need to be in the forefront of development, evaluation, dissemination, and advocacy. New and important issues abound.

This emerging field requires development of a broad cadre of investigators who understand the principles and practice of item improvement, quantitative item calibration, and combining of items into instruments using IRT and CAT. It requires validation of new instruments in different

From the Department of Medicine, Stanford University School of Medicine, Stanford, California; National Institutes of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), Bethesda, Maryland, USA; Medical School Charité, University Medicine Berlin, Berlin, Germany; Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, Illinois; Department of Medicine, University of Michigan School of Medicine, Ann Arbor, Michigan; Department of Pediatrics, University of Cincinnati School of Medicine, Cincinnati, Ohio, USA.

Supported under the PROMIS-Initiative funded by NIAMS, with multiple grants to the listed institutions and others to support development of improved patient-reported outcomes items, instruments, and applications.

J.F. Fries, MD, Professor of Medicine (Emeritus), Department of Medicine, Stanford University School of Medicine; J. Witter, MD, PhD, Medical Officer, NIAMS; M. Rose, MD, Chair, Psychosomatic Medicine, Medical School Charité, University Medicine Berlin; D. Cella, PhD, Chair, Medical Social Sciences, Department of Medical Social Sciences, Feinberg School of Medicine; D. Khanna, MD, Associate Professor of Medicine, Department of Medicine, University of Michigan; E. Morgan-DeWitt, MD, Assistant Professor of Pediatrics, Department of Pediatrics, University of Cincinnati School of Medicine.

Address correspondence to Dr. Cella; E-mail: d-cella@northwestern.edu

diseases, domains, and languages. All existing outcome assessment instruments intended for broad use require re-assessment and improvement, and new measures need to be developed to a higher standard. Details of specific studies and populations are provided in cited articles[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]. Adoption of efficient, precise, and standardized instruments in clinical trials and in longitudinal studies is much needed. Investigators with the requisite interests and skill sets to transform clinical research will be in short supply during the transition.

Three concepts that frame the issues are the Patient-reported Outcomes Measurement Information System (PROMIS) domain framework (an organized set of uni-dimensional outcome domains); the potential of IRT to inform the development, scoring, and application of improved items measuring these domains; and the potential of CAT to achieve improved measurement with fewer questions.

Domain Framework

With IRT, a framework of self-reported health is based on the concept of outcome "domains" rather than disease-based classifications. The initial domains adopted by the World Health Organization are physical health, mental health, and social health. For PROMIS[10], physical health logically divides into domains of physical function and physical symptoms. Physical function entails assessing function below the population mean, sometimes termed disability, as well as function above the mean, sometimes termed fitness, wellness, or positive health. Physical function may be subdivided into function of the lower extremities (mobility) and function of the upper extremities (dexterity). Similarly, mental health and social health can be broken down into component domains. Figure 1 shows a simplified PROMIS framework[16]; PROMIS actually has over 40 domains under study. All PROMIS items and data have been reviewed by individual institutional human subjects research protection programs, and each participant gave written informed consent.

The PROMIS domain framework is intended to be loosely hierarchical, indicating that subordinate domains illustrated on the right side of the framework can be aggregated into the more general domains to the left, but the reverse is not the case. For example, walking is a subdomain of mobility, which in turn is a subdomain of physical function. For unidimensional IRT models to be applied, all of the items in a domain must represent the same underlying concept. In addition, items should not be redundant with one another; rather they should each measure different aspects of the same underlying concept, such as pain or physical function.

The domain of physical function replaces the historic term disability. Disability is traditionally expressed in decrements below normal; its measurement has been limited by ceiling effects and a logical conundrum: If people have a disability level of zero, does this mean that they cannot improve their health? If a patient with rheumatoid arthritis (RA) improves her physical function to above average (e.g., by medication and exercise), does this not mean that she should be assessed as having a better response than a patient who improves only to the population norm? Physical function assesses both increments and decrements. Disability assesses only decrements. Therefore, physical function scores are now presented as T scores where the population mean is set to zero and each unit above or below the mean is 1 SD and is represented by 10 points on the scale (Appendix 1)[17]. We have to learn to use physical function scores, unfamiliar as they seem at the beginning, if we are going to be able to study improvements in human health.

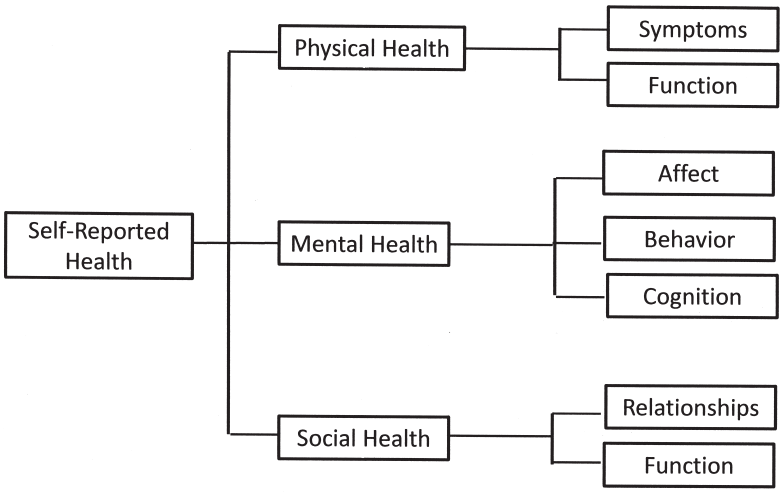


Figure 1. The PROMIS hierarchy of domains. A simplified diagram of the first 3 levels of the hierarchy. Patient-reported Outcomes Measurement Information System (PROMIS®) domain framework[16]. © 2011 PROMIS Health Organization and PROMIS Cooperative Group. Adapted with permission.

Item Response Theory

IRT[3,4] focuses the approach to PRO assessment upon the individual item's representation of the underlying metric, rather than at the level of the outcome assessment instrument such as the HAQ[1] or the PF-10[2]. IRT-based instruments are constructed from items selected from the item bank representing a specific domain of health. The information content of individual items and the correlations of individual items with a latent trait (such as physical function) may be quantitatively "calibrated" using IRT. Attributes such as clarity, ease of translation, and importance to the patient may be qualitatively evaluated and used to winnow item sets into fewer, better items. After discarding unclear, obscure, trivial, and redundant items, improved instruments may readily be constructed for testing and validation.

The goal of a PROMIS item bank is that it includes a large collection of items, developed and evaluated according to PROMIS standards (www.nihpromis.org/Documents/PROMIS_Standards_050212.pdf). An item is a question with a range of response options, such as "Are you able to walk a block? (without difficulty, with a little difficulty, with some difficulty, with much difficulty, or unable to do)". An item has a "stem," a "time frame," a "context," and a set of "response options."

Items are calibrated by measuring the degree of difficulty of the item and the degree with which the item distributes among the several response options provided. Items may be substituted for other items that have the same calibration scores. Although calibration models are complex, and full discussion is beyond the scope of this article, PROMIS investigators made a decision to use the graded response model as the primary approach, with analysis of alternative models, such as Rasch models, to ensure consistency of results. We have found that across these various models, there is little difference in results obtained, suggesting they are often interchangeable, or at least highly consistent in the end result[11].

A major finding from work with calibrated items is that outcome instruments consistently perform best across a relatively narrow part of their theoretical range. This occurs because there are few if any items included in the scale that represent extremely low function, perhaps as in a patient in a nursing home, or items that can well assess a trained athlete with function far above average. Thus, the scale usually has had major "floor" limitations in those with very poor function, and major "ceiling" limitations affecting assessment of those with extremely high levels of functioning. The remedy is to develop new items that work well at the floor and other new items that work well at the ceiling, and then to calibrate the new items in subject populations with very high or very low functional abilities.

Computerized Adaptive Testing

CAT[5,6] provides a means to dynamically administer items from a calibrated bank, emphasizing items that provide the greatest information increment given a subject's prior responses. Those items that tend to be most frequently selected in a test population may also be aggregated to form a static short form tailored to a portion of the severity range represented by the test population.

PROMIS items, selected and improved from extensive prior experience and literature on nearly 10,000 items across multiple health concepts, were tested in over 21,000 participants. The data collected in early PROMIS years are available to other researchers upon request (see www.NIHPROMIS.org). Item characteristics can thereby be tested and re-evaluated using various IRT models, including Rasch models.

A typical clinical study determines for each subject a baseline value and a final value, and the study seeks to determine a change score and compare it with a change score from another arm of the study. The change score is an estimate of the true change over time; the true change score is a "latent trait," which cannot be directly observed because there are error terms around the baseline score and about the final score. If the error terms are reduced, the estimate of the true change becomes more precise. IRT and CAT allow closer estimates of the true change and this improves study power.

There are 3 general designs in which CAT applications may construct endpoints for a clinical study. First, CAT can be administered to each individual on each occasion. This means that the subject will seldom get the same items at the end of the study as they did at the beginning, but the precision of each value will be better than by other designs because the best items for that subject at that time were always being asked. If there is a long delay between baseline and final measures (e.g., years), and physical function levels may have shifted because of the passage of time, this may well be the preferred design.

A second design would be to use CAT at baseline to get a score for each person, and at the final observation to administer exactly the same items previously selected by CAT. Here, each subject gets his or her own items at both administrations. Such a design may be preferred in a clinical trial, because it is somewhat simpler than a design that always uses CAT, and because the comfort level of the investigators may be higher if the same items are used for the same person throughout the study. A potential limitation is, for example, if a subject makes dramatic improvement, the items selected at baseline may constrain measurement of the improvement by imposing a ceiling rather than allowing the CAT algorithm to select items at higher levels of the construct.

A third CAT design may use a CAT on a sample of patients similar to those expected in the clinical study.

Frequently chosen items can then be compiled into a static short form that can then be administered to all patients throughout the study. In this case, all participants complete the same form throughout the study, with a form that has been customized in advance to be preferentially targeted (i.e., responsive) to the level of the patient population. This approach can be very close to the efficiency of the first 2 designs in the case of physical function, if a 20-item instrument is used. With other item banks, the form length is typically fewer than 10 items. In an extension of this design, CAT-tailored short forms each designed to assess one-fifth of a severity distribution, for example, might be an option to reduce floor and ceiling effects. Static forms can never be quite as precise, however, as when a CAT is derived for each subject and they remain more likely to have floor and ceiling problems than do the first and second design options.

IRT-based Instruments

In some domains (such as physical function), IRT-based instruments already are known to perform significantly better, require smaller sample sizes, have greater reliability, have lower questionnaire burden, and result in greater responsiveness than previously available instruments, both in randomized clinical trials and in longitudinal observational studies[7]. Ongoing efforts, to which interested persons are invited, include extension of these domains to additional domains and new clinical populations, major reduction in floor and ceiling effects, linkages to cost and drug toxicity domains, and adoption in additional clinical trials, observational studies, and individual patient care[7,8,9,10,11,12].

MATERIALS AND METHODS

The PROMIS initiative is a very large, multiinstitutional US National Institutes of Health (NIH) Common Fund program intended to develop and make available improved PRO instruments to broadly enhance the quality of clinical science. PROMIS is currently in its second phase of NIH funding, and many of the investigators using IRT and CAT methodology have been associated with PROMIS since its inception. In 5 domains (physical function, pain, fatigue, emotional distress, and social support), all English language items from existing instruments (about 8000) have been identified, catalogued, and categorized by PROMIS investigators, and redundant items and those not focused on the latent trait eliminated.

Remaining items were improved for clarity, ease of translation, and importance to patients, with attention to the item stem, item content, time frame, and response options. Qualitative evaluation used focus groups, patient surveys, and modified Delphi techniques. Remaining items, about 700, were evaluated quantitatively using IRT, in population-based and clinical disease groups including over 21,000 subjects. The strongest items, balanced for content in each domain, were assembled into short-form static instruments of 10 to 20 items, and CAT applications using up to 10 items were selected dynamically for each patient.

We sought to compare instruments to determine the magnitude of improvement with IRT-based and CAT-based instruments as compared with legacy instruments. We compared the most widely used legacy instruments (Legacy HAQ and Legacy PF-10); these same instruments after improvement in item stems, response options, context, and time frame (Item-Improved HAQ and PF-10); and the PROMIS 20-item PF Short Form derived from a 154-item item bank using IRT techniques. Finally, we assessed a 10-item CAT based on the same item bank. Order of instrument administration was randomly assigned. The primary outcomes were effect sizes and sample sizes required for 80% power to detect an alpha level of 5%. We set scores on all instruments to range from 0–100, where 100 is the worst function. We studied 451 patients with RA at baseline and 1 year. The endpoint tested was average progression of RA over 1 year. All patients received the same items, but the order of presentation was randomly assigned.

RESULTS

All instruments could detect average change at 1 year with $p < 0.05$, ranging from $p < 0.01$ to $p > 0.04$. All changes were in the same direction and of similar magnitude. The minimum detectable differences ranged from 1.14% and 1.24% with the Item-improved HAQ and the PROMIS PF-20 and 2.43% with the Legacy PF-10. Differences in sample sizes required were dramatic: about 100 subjects per arm when using the 2 best-performing instruments and 427 subjects per arm when using the PF-10.

Appendix 1 shows comparative information content across the severity spectrum for 6 of the instruments tested. This "boat diagram" summarizes data on instruments being compared, and will reward careful study. The horizontal axis notes t scores where zero is the population mean and each number above or below the mean indicates the number of SD above (to the right) or below the population mean (to the left). The vertical axis indicates reliability in terms of the standard error of measurement. A standard error of 3.2 corresponds to a reliability of 0.90 and a standard error of 2.2 represents a reliability of 0.95. A summary measure sometimes useful to compare the power of an instrument is the number of SD range of values covered with a reliability (precision) of 0.90 or better; this measure approximates an area over the curve. The legacy PF-10 covers 2.4 SD with a reliability of 0.90 or better, but the static PROMIS PF-20 covers 4.8 SD[7]. The PROMIS PF-10-item CAT covers 6.3 SD and demonstrates marked superiority at the ceiling.

Thus, the best instruments will have the lowest curves with the broadest coverage of extremes in the population. The PROMIS PF-20 is the most effective static instrument by these criteria, although the curves cross those of the HAQ and the Item-Improved HAQ about 3 SD below the mean. Static forms with 20 items outperform those with 10 items. The HAQ outperforms the PF-10 in more disabled populations; the PF-10 outperforms the HAQ in normal populations. Static forms with 5 response options outperform those with 3 or 4. The PROMIS CAT, although limited to 10 or fewer items, outperforms all static instruments tested. Only an impossibly ponderous instrument containing all the items in an item bank would be expected, on theoretical grounds, to outperform the CAT. Better floor and ceiling items will further extend the range of either tailored static forms or CAT.

PROMIS validation studies are now completed or under way in dozens of domains, as are translations and additional studies across a range of chronic disease. PROMIS items

and instruments are in the public domain. As expected, results are strongly positive. Longitudinal studies of sensitivity to change suggest that the effect size is increased by 0.1 to 0.2 by the new instruments as compared with the old. Three randomized trials have been reported, and multiple observational studies have been completed and submitted for publication; many more are in process. Proof of concept randomized trials involving the HAQ and other instruments have been completed in over 1000 patients. Two randomized clinical trials of mode of administration (mailed, Internet, hand-held device, telephone) confirm that bias is not introduced by using any of the major modes of item administration. The PROMIS Assessment Center, an online PROMIS clinical study management system, is essentially in the public domain, as are the PROMIS item banks and short forms. Assessment Center currently manages over 50 clinical trials.

DISCUSSION

Advances in PRO measurement can permit more statistically powerful and efficient clinical research. We suggest their use should be expanded as rapidly as possible. Those involved (e.g., US Food and Drug Administration, industry, NIH, academia, World Health Organization, US Department of Health and Human Services, and Patient-centered Outcomes Research Institute) may contribute to the improvement of the field, because they have an interest in improved outcome assessment. Additional domains, diseases, patient populations, languages, validations, and collaborations are needed. The role of PROMIS is as a framework and a resource.

This field is evolving rapidly toward an item-based, not instrument-based outcomes assessment. Here, for example, a core PROMIS PF item bank, with calibrated items, translated into several languages, in the public domain, augmented as necessary, growing over time, and open to all, will accelerate new instrument development and use of various CAT approaches, all based upon different configurations of the core PF item bank.

The initial PROMIS item banks, for the most part, have underrepresentation of items that measure the health construct at the extremes, and this gives rise to floor and ceiling effects, which reduce general applicability and decrease study power. These are being effectively eliminated in the physical function domain, and soon will be in others. A practical limitation to the field has been the need to develop more efficient ways to collect data from multiple simultaneous inputs and remote inputs. These are being approached by development of Microsoft iPAD app and wireless communication. Patient care applications, long an ideal, will become increasingly attractive with instantly available scores and easy access from any location[13,14,15].

We recommend the PROMIS PF-20 as the best instrument to replace applications where an investigator would typically use the traditional 20-item HAQ. It is more precise than the Item-improved HAQ, and compatible with the shorter PROMIS PF CAT. PROMIS improvements in the HAQ have been evolutionary rather than revolutionary. They are similar to prior HAQ revisions over the years as when car doors lost their push buttons and people generally stopped taking tub baths. We believe firmly that the many hundreds of validation studies in scores of languages for the HAQ apply to the PROMIS instruments as well; these differ principally in the use of 5 response options rather than 4 and the explicit use of the present tense. PROMIS items have many advantages over legacy items, such as IRT, CAT, translation, cultural adaptation, consistency, clarity, patient input, efficiency, and permitting smaller study sample sizes. They have no disadvantages. But the CAT options, when fully functional, will greatly outperform the PROMIS PF-20.

Future Directions in Rheumatology

The PROMIS mission is to use measurement science to create a state-of-the-art assessment system for self-reported health to advance PRO measurement in clinical research and day-to-day practice. Similar to other PRO, this will facilitate the incorporation of the patient's voice into clinical trials and clinical practice.

Using PROMIS in Clinical Practice

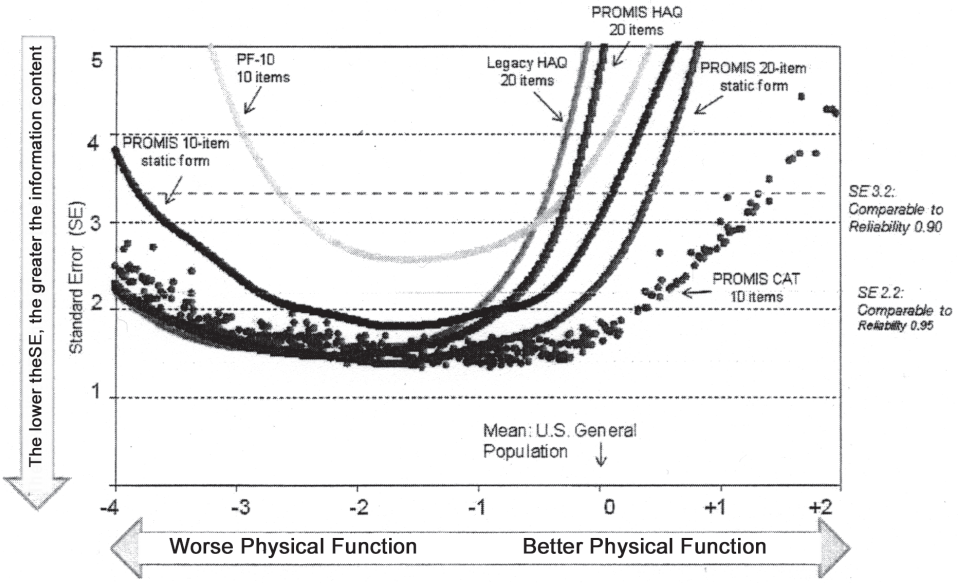
PROMIS has many advantages. It allows administering item banks in the waiting room on a personal computer or in paper-and-pencil version and have instant scoring (using CAT). It can be calibrated to population norms. It is ready to share with the patient at point of care. As an example, the feasibility of 11 PROMIS item banks was recently assessed in a single-center, observational study in patients with systemic sclerosis[8]. The average number of items completed for each CAT-administered item bank ranged from 5 to 8 (69 CAT items per patient), and the average time to complete each CAT-administered item bank ranged from 48 s to 1.9 min per patient (average time 11.9 min per patient for 11 banks).

PROMIS has developed item banks that are relevant to rheumatology, can be "customized," and are currently freely available. The item banks provide tremendous flexibility for creation of fixed-length short-forms or CAT administration. This quick assessment can generate a patient report to monitor health over time.

REFERENCES

1. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
2. Ware JE Jr, Kosinski M, Keller SD. A 12-item Short Form Health Survey. Med Care 1996;34:220-33.
3. Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, Ware JE Jr. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther 2009;11:R191.

APPENDIX 1. Comparison of information content between 6 instruments. The most precise instruments have curves that are below and broader than the less-reliable instruments. Reliability at the "floor" is represented by the left side of the curves and reliability at the "ceiling" is represented by the right side of the curves. The 10-item PROMIS CAT covers 6.4 population SD and has much greater precision in normal populations (less ceiling effect). Reprinted from Fries, et al. J Rheumatol 2009;36:2061-6. PROMIS: Patient-reported Outcomes Measurement Information System; HAQ: Health Assessment Questionnaire; PF: physical function (domain); CAT: computerized adaptive testing.



4. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63:1179-94.
5. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061-6.
6. Fries JF, Krishnan E. What constitutes progress in assessing patient outcomes? J Clin Epidemiol 2009;62:779-80.
7. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS disability scales with item response theory. Arthritis Res Ther 2011;13:R147.
8. Khanna D, Hays RD, Maranlan P, Rothrock N, Cella D, Gershon R, et al. Feasibility and evaluation of the construct validity of PROMIS and "legacy" instruments in an academic scleroderma clinic. Value Health 2012;15:128-34.
9. Khanna D, Krishnan E, DeWitt EM, Khanna PP, Spiegel B, Hays RD. The future of measuring patient-reported outcomes in rheumatology. Arthritis Care Res 2011;63 Suppl 11:S486-90.
10. Patient-Reported Outcomes Measurement Information System (PROMIS). [Internet. Accessed July 22, 2013.] Available from: www.nihpromis.org
11. Riley WT, Rothrock N, Bruce B, Christodolou C, Cook K, Hahn EA, et al. Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: further evaluation of content validity in IRT-derived item banks. Qual Life Res 2010;19:1311-21.
12. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary function item bank supported the expected advantages of the patient-reported outcomes measurement information system (PROMIS). J Clin Epidemiol 2008;6:17-33.
13. Fries JF, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and internet administration. J Rheumatol 2011;38:1759-64.
14. Fries JF, Krishnan E, Bruce B. Items, instruments, crosswalks, and PROMIS. J Rheumatol 2009;36:1150-7.
15. Rose M, Bezjak A. Logistics of collecting patient-reported outcomes (PROs) in clinical practice: an overview and practical examples. Qual Life Res 2009;18:125-36.
16. PROMIS domains. Domain hierarchy framework. PROMIS Health Organization and the PRO MIS Cooperative Group; 2008. [Internet. Accessed July 22, 2013]. Available from: www.nihpromis.org/Documents/PROMIS_Full_Framework.pdf
17. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061-6.