HHS Public Access

Author manuscript

J Clin Epidemiol. Author manuscript; available in PMC 2015 June 14.

Published in final edited form as:

J Clin Epidemiol. 2014 May ; 67(5): 516–526. doi:10.1016/j.jclinepi.2013.10.024.

The PROMIS Physical Function Item Bank Was Calibrated to a Standardized Metric and Shown to Improve Measurement Efficiency

Matthias Rose[1,2], **Jakob B. Bjorner**[3,4,5], **Barbara Gandek**[1], **Bonnie Bruce**[6], **James F. Fries**[6], and **John E. Ware Jr.**[1,7]

[1]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

[2]Department of Psychosomatic Medicine, Medical School, Charité Universitätsmedizin Berlin, Germany

[3]National Research Centre for the Working Environment, Copenhagen, Denmark

[4]i3 QualityMetric Incorporated, Lincoln, RI, USA

[5]Institute of Public Health, University of Copenhagen, Denmark

[6]Stanford University School of Medicine, Palo Alto, California, USA

[7]John Ware Research Group, Worcester, MA, USA

Abstract

Objective—To document the development and psychometric evaluation of the Patient-Reported Outcomes Measurement Information System (PROMIS) Physical Function (PF) item bank and static instruments.

Study Design and Setting—Items were evaluated using qualitative and quantitative methods. 16,065 adults answered item subsets (n>2,200/item) on the Internet, with over-sampling of the chronically ill. Classical test and item response theory (IRT) methods were used to evaluate 149 PROMIS PF items plus 10 SF-36 and 20 HAQ-DI items. A graded response model was used to estimate item parameters, which were normed to a mean of 50 (SD=10) in a US general population sample.

Results—The final bank consists of 124 PROMIS items covering upper, central, and lower extremity functions and IADL. In simulations, a 10-item Computerized Adaptive Test (CAT) eliminated floor and decreased ceiling effects, achieving higher measurement precision than any comparable-length static tool across four standard deviations of the measurement range. Improved

Corresponding Author: Matthias Rose, MD, PhD, Matthias.Rose@charite.de.

psychometric properties transferred to the CAT's superior ability to identify differences between age and disease groups.

Conclusion—The item bank provides a common metric and can improve the measurement of PF by facilitating the standardization of PRO measures and implementation of CATs for more efficient PF assessments over a larger range.

Keywords

Item Response Theory; Computer Adaptive Test; physical function; health status; questionnaire

1. Introduction

Measurement of patient-reported outcomes (PRO) in clinical studies has steadily increased in frequency, as has its importance in evaluating therapies and developing treatment plans. The plethora of outcomes tools available today allows for increasing specificity of measurement over a wide range of domains. However, most widely-used PRO tools have well-described shortcomings, including high respondent burden and lack of measurement precision. Moreover, results from different instruments can be hard to compare, which limits the interpretability of PRO data.

To address these shortcomings, the National Institutes of Health funded an initiative to build a comprehensive Patient-Reported Outcomes Measurement Information System (PROMIS) [1;2]. PROMIS uses Item Response Theory (IRT) and Computerized Adaptive Tests (CATs), which are believed to be promising solutions to the most important limitations of current measurement tools. An IRT item bank consists of a set of items measuring the same construct and parameters that describe the items' measurement properties [3]. Item banks provide the foundation for CATs, which make it possible to administer the most informative items to an individual respondent [4]. Thus, higher precision can be achieved, while respondent burden can be reduced [5–7].

One aim of the PROMIS initiative was to build an improved item bank for the Physical Function construct, which has been evaluated using IRT methods for more than a decade [8–12]. Items covering a wide range of functioning, from self-care to strenuous activities, have been calibrated using IRT models, and some of the first CATs were developed for Physical Function [13;14]. We presented results from a rigorous evaluation of IRT methods in preparation for development of the PROMIS Physical Function item bank earlier in this journal. [6]. The current paper reports on the development and initial psychometric evaluation of the PROMIS Physical Function item bank.

2. Methods

Item bank development and evaluation followed the general PROMIS approach, described in detail elsewhere [2;6;15;16]. Issues specific to the PROMIS Physical Function item bank are discussed below.

2.1 Item Bank Development and Qualitative Review

PROMIS aimed to develop a generic Physical Function item bank that could be used across diseases and different levels of ability. Four sub-domains were defined: instrumental activities of daily living (IADL), mobility or lower extremity function, back and neck (central) function, and upper extremity function. The qualitative work to develop the Physical Function item bank has been described in detail [17]. In short, 1,728 items from 165 instruments were reviewed and 1,560 items were eliminated as redundant, condition-specific, vague or unrelated to the domain. Most remaining items were rewritten to minimize variation in item attribution and response scales. As in many existing Physical Function measures, items used the present tense. Items that were primarily determined by the respondent's functioning omitted health or disease attribution, because difficulty in performing these tasks was considered to be due to health problems or disability. Tasks that were strenuous or included social interaction might be constrained by non-health related factors, so these included health attribution. All items went through additional expert review and patient assessment.

The field test included 168 Physical Function bank items, one global Physical Function item, and 30 items from two legacy tools (20 Health Assessment Questionnaire (HAQ-DI)[18] and 10 SF-36 Health Survey Physical Functioning (PF-10) [19] items). Twenty of the 168 items are not analyzed here, including 5 items about device use, 7 items about task performance, 7 items that were not specific to Physical Function, and 1 item that we could not obtain permission to include in the final bank.

2.2 Field Testing

21,133 adults participated in the PROMIS Wave 1 data collection, 16,065 of whom answered two or more Physical Function items. Of these, 14,777 were enrolled via a YouGov/Polimetrix Internet portal [20], 54.7% from the general population and 45.3% self-identified with specific diseases. Another 1,288 enrolled at four PROMIS clinical sites. All participants answered 10 PROMIS global items [21] and questions on clinical conditions and sociodemographics. A sub-sample (Form C) completed the HAQ-DI and PF-10.

The 168 PROMIS items were administered in two different designs. In the 'full bank' design, one sub-sample (Form C) answered 112 items, while another sub-sample (Form G) answered the remaining 56 items. This full bank design allowed for analysis of the item covariance matrix without using imputation methods. Within the 'block design' (Forms H-W), 16 sub-samples answered different subsets of items from all PROMIS item banks, including 21 Physical Function items each. This balanced incomplete block design allowed for simultaneous IRT-based estimation of item parameters, i.e., blocks of at least 7 items from each domain were administered in two independent samples. Items were administered in a fixed order in both designs.

2.3 Analysis

2.3.1 Data Preparation and Skewness—If fewer than three participants endorsed a response option for an item, we collapsed that response option with the adjacent response option. The full set of response options will be used in future item administrations, but

collapsed response options were used for score estimation. 248 participants (1.54%) were excluded because they had response patterns or response times indicating insufficient attention or had too much missing data [22], resulting in a sample of 15,817. Skewness was used to indicate poor fit between the health of the sample and the level of health measured by an item.

2.3.2 Unidimensionality and Local Independence—To ensure that items were measuring Physical Function, items correlating <.50 with the global Physical Function item were excluded. Confirmatory factor analysis (CFA) was used to explore the interrelationship of the four a priori defined Physical Function sub-domains and to determine if a sufficiently unidimensional Physical Function construct [23;24] could be obtained, using the Mplus™ with an WLSMV estimator [25]. Items with factor loadings below .70 in the final CFA were eliminated [26]. Studies from Wolfe [12] and our previous work [6] supported a two factor solution. Accordingly, we tested three alternative ways to define a two factor solution: (1) Fine (hand activities) vs. Gross Motor Activities; (2) Upper (hand or arm activities) vs. Back/Neck and Lower Extremity; and (3) Musculoskeletal (all Upper Extremity items) vs. Cardiopulmonary Demanding Tasks.

To test for local independence [15], we analyzed residual correlations using Mplus™ [27]. If a pair of items had a residual correlation of .25 or more we eliminated the item that had a higher accumulated residual correlation with the remaining items [28].

2.3.3 Differential Item Functioning—Tests of differential item functioning (DIF) [29] were used to identify systematic error due to group bias (independent variables were gender, age, education, and disease), using an ordinal logistic regression model in which the item response was regressed on the total sum score of all items and each independent variable. A significant effect of the independent variable on the item response indicated uniform DIF, while a significant interaction effect (between the independent variable and sum score) indicated non-uniform DIF. The magnitude of DIF was evaluated with the coefficient of determination R^2 as described by Nagelkerke[30]. An increase in combined $\Delta R^2 > 0.03$ indicated noticeable DIF. DIF for age, gender, and education were evaluated in the 'full bank' data. DIF for disease (musculoskeletal, cardiopulmonary, mental) was evaluated twice for each item within the 'block design' data.

2.3.4 Monotonicity—Item response curves (IRC) were examined using the program TestGraf [31], applying a non-parametric kernel-smoothing technique. Each response option curve should have only one clear maximum that is well separated from the maximum of other curves.

2.3.5 Item Parameter Estimation and Item Fit—Item parameters were estimated using a Graded Response Model (GRM) [3] with Multilog Version 7. Parameters for the PROMIS item bank and global item were estimated first. Item fit statistics were calculated based on algorithms published earlier[32], using the SAS macro IRTFIT [33]. We report S-G^2 values (a likelihood ratio G^2 statistic), which quantifies the difference between expected and observed frequencies of item category responses for various levels of scores. Non-fitting items were identified if test statistics were significant in at least two of the three (one full-bank and two

block-design) sub-samples. Item parameters then were re-estimated excluding the non-fitting items and fit tests were re-evaluated. This procedure was repeated until all items in the model fitted. Item parameters then were fixed for the fitting items, and item parameters were estimated for the non-fitting items.

Once item parameters were established for the PROMIS items, parameters were estimated for legacy HAQ-DI and PF-10 items, holding the PROMIS item parameters as fixed. For this estimation, PROMIS items with similar content as legacy items were excluded. To counter skewness for the legacy items, ARAMIS data (phase 48.1 http://aramis.stanford.edu/ [18]) with the HAQ-DI a

2.3.6 Population-Based T-Score Transformation—IRT-calibrated scores were transformed to have a mean of 50 and standard deviation of 10 in the U.S. general population, as described elsewhere [20]. All item parameters were centered based on the mean and standard deviation of this scale-setting sample. Higher scores indicate better physical function.

2.3.7 Analysis of Item Information Functions and Reliability—Item information functions (IIF) were calculated from the IRT model [34]. An IIF describes each item's contribution to overall test precision, and their sum defines the ideal precision of the test at a given level of the latent trait (Θ theta), allowing for estimation of the expected standard error. For samples with an IRT score standard deviation of σ=10, a standard error of 2.3 is comparable to an internal consistency of α=0.95.

2.3.8 Static Form Development—To demonstrate a potential use of the PROMIS item banks, we constructed 10- and 20-item static forms that covered similar content as legacy tools, included all four sub-domains, balanced items measuring 'ability' and 'limitations', covered a wide measurement range, and provided good measurement properties. Thus, the static forms were constructed based on both content and psychometric considerations. A third static form with 5 items that excluded upper extremity items also was tested. Each shorter static form contains a subset of items from the preceding longer form.

2.3.9 Simulation Studies—Simulation studies were performed to describe properties of the item bank, static forms, and potential CATs. To cover the range in which most patients would score, we simulated the answers of 1,000 simulees having a normal distribution with a mean of 40 and a SD of 20. We also simulated a 10-item CAT for a general population sample (mean=50, SD=10) and a potential clinical sample (mean=30, SD=10).

2.3.10 Validity Testing—Construct validity was evaluated by correlating scores for the item bank and static forms with scores for two legacy measures (HAQ-DI and PF-10). We also used the method of known-groups validity and conducted analyses of variance to determine how well PROMIS and legacy measures distinguished among groups varying in self-reported health, age, and number of chronic conditions. Relative validity (RV) coefficients were computed for each measure in each test by computing the ratio of pair-wise F-statistics, with the F value of the item bank as the denominator. The RV coefficient indicates in proportional terms how valid a scale is relative to the item bank.

3. Results

3.1 Sample

Fifty-two percent were female. Age ranged from 18–95 with a mean of 54. Eighty-two percent were white, 9% black and 8% multi-racial; 9% Hispanic or Latino. Education ranged from less than high school (3%) to an advanced degree (19%), with 24% having a college degree, 38% some college, and 16% a high school diploma. The majority reported at least one chronic condition, but most reported no limitations in carrying out daily activities (Table A-1).

3.2 Skewness

For 50% of items, at least 75% of subjects endorsed the least difficult response option, and almost 90% endorsed the two least difficult options. Two highly skewed items (skewness < −7.25; ‘wash face’, ‘open and close mouth’) were excluded. The least difficult response options for 11 items were collapsed because fewer than three respondents endorsed the least difficult category.

3.3 Unidimensionality and Local Independence

Two items did not show a sufficient correlation (r<.50) with the global Physical Function item (‘open new or tight jar’, ‘turn head side to side’), and were excluded.

In the four factor CFA, all factors had very high correlations (r=0.89–0.97), supporting a more parsimonious solution. All three alternative two factor solutions for ‘full bank’ samples produced very similar results and showed high two factor correlations (Form C r>. 90/Form G r>.75) (Table A-2). Fit indices changed minimally for a one factor solution. Form C data showed reasonably good fit for the one and two factor solutions (RMSEA 0.084 vs 0.088). Form G data showed a less favorable fit, but it was similar for both solutions (RMSEA 0.143 vs 0.133). Some ‘block design’ analyses suggested that a two factor solution provided a slightly better fit, but in all samples the two factors were highly correlated (mean r=0.87, median r=0.88). Even in the worst fitting one factor solutions, fit indices were still in the range frequently seen in WLMSV estimates for health questionnaires [35]. Thus, we pursued a more parsimonious one factor solution, as this was more practical and the resulting item bank enables a wider range of measurement. One item with a factor loading <.70 in the one factor solution (‘turn head’) was excluded. No residual correlations of the remaining items were above 0.25, so no items were excluded for local dependence.

3.4 Differential Item Functioning

Out of 429 tests in the ‘full bank’ sample (143 remaining items and three socio-demographic variables), only 12 tests showed DIF (4 age, 7 gender, 1 education). Overall, no clear pattern was observed. In ‘block design’ data, ten items showed DIF for disease in at least one of the two tested samples for each item. Three items demonstrated particularly meaningful (>0.05) differences, two of which were in Form L (predominantly patients with musculoskeletal conditions), which is highly dominated by dexterity items. Figure A-1 shows that patients

with musculoskeletal disease indicated relatively less impairment in standing up or walking than patients with cardiovascular disease at a given Θ-level.

Given the multitude of DIF tests, the number of items identified as showing DIF was small. Thus, we followed the PROMIS strategy of retaining these items to allow further analysis of their impact.

3.5 Monotonicity

No item showed violations to the monotonicity assumption.

3.6 Item Parameter Estimation and Item Fit

Item parameters were estimated for the remaining 144 PROMIS items (143 item bank and one global item). Seven iterations of item fit tests were performed until a final IRT model included only items with no misfit. Most non-fitting items either asked about strenuous or very easy activities. When experimental items (used to evaluate different item stems) and misfitting items were removed, parameters for 124 PROMIS items (plus one global item) remained. We additionally estimated parameters for the HAQ-DI and PF-10.

3.7 Item Bank Properties

Across all items, discrimination parameters (slopes) were high with a mean of 3.17 (± 0.70). The mean maximum item information was 2.68, with a range from 0.72 (‘open jars’) to 5.58 (‘chores like vacuuming, yard work’). Most items provide the best information around a Θ-value of 30 (2 standard deviations below the population mean), but maximum information ranged from a Θ of 10 (‘lift cup to mouth’) to 65 (‘run ten miles’) Items with the highest information have their maximum around a Θ-value of 40. Table 1 illustrates these properties for the items included in the static forms.

3.8 Static Form Development

Figure 1 demonstrates the precision that can be expected in comparison with the criterion standard (entire item bank) and a simulated 10-item CAT. The 20-item static form matched the expected precision for a 10-item CAT but had a more restricted range. Omitting items about upper extremity functions in the 5 item static form illustrates the loss of measurement precision and an increased floor problem.

3.9 Simulation Studies

Simulation studies showed that an IRT-scored SF-36 Physical Functioning scale provided very good measurement properties (SE <2.3) around a range of two SD below the U.S. general population mean, and that an IRT scored HAQ-DI provided very good measurement properties around a range of 4 SD below the mean. However, the same measurement precision could be obtained over a substantially larger measurement range if a 10-item CAT was applied (Figure 1).

3.10 Validity Testing

Within the Form C data, a 10-item CAT correlated r=.98 with the IRT score from all bank items. The CAT also correlated strongly with the PROMIS static forms (static20/10/5, r=.85/.88/.90) as well as with the SF-36 PF-10 (r=.86), and lower with the HAQ-DI (r=−.67). PROMIS static forms correlated highly among each other (r>.90), and had similar high correlations with both legacy tools (PF-10 with static20/10/5 r=.86/.91/.91; HAQ-DI with static20/10/5 r=−.91/−.86/−.77).

All tools discriminated across groups differing in self-reported general health, age, and number of chronic conditions (Table 2). In almost all instances, the full PROMIS item bank and CAT showed a higher relative validity than the PROMIS static forms and legacy instruments. The PROMIS static forms showed higher F-values than legacy tools of same length.

4. Discussion

PROMIS aims to make a major contribution to improved measurement of patient-reported outcomes. The development of the Physical Function item bank is one part of the project.

An important contribution of PROMIS is its extensive qualitative work. The literature search used to build the Physical Function item bank [17] defines this construct based on the body of instruments that have emerged over the last three decades. The resulting item bank contributes to the long-term goal of substituting an instrument-defined measurement system with a construct-defined measurement system, where different tools can be scored on one common metric.

As in our previous work [6], we showed in simulation studies that CATs are likely to outperform static tools of the same or longer length in measurement precision and range, as well as discriminant validity. Based on these simulation studies, it can be expected that a 10-item PROMIS Physical Function CAT will be able to measure Physical Function with high precision (comparable to a reliability of .95) over a range of more than six standard deviations. Reducing floor and ceiling effects addresses a serious shortcoming of most disease-specific tools as many chronically ill patients also experience periods with normal functioning (i.e. $\Theta \approx 50$ equivalent to a HAQ-DI-score$\approx$0.07). While real CAT applications need to confirm this finding, this is an important proof of concept. PROMIS measures also correlated highly with established Physical Function tools, demonstrating construct validity [36].

However, in addition to these encouraging findings, this research raised a number of noteworthy issues that need to be addressed.

4.1 Conceptual Issues

A major issue is the dimensionality of Physical Function, which has been explored at length. Previous research has supported a two dimensional (upper versus non-upper) approach[6;12;37], although Martin also found that a one factor model was more responsive to clinical outcomes[37]. Raczek[11] and Hays[38] also showed that mobility, self-care, and back and

neck functions fit a one factor IRT model reasonably well. Our analysis demonstrated that upper, central, lower body and IADL items can be combined, reflecting the assumption that each item measures an underlying Physical Function construct. This replicates what has been successful for a classic tool such as the HAQ-DI. However, for some specific diseases, it is likely that some activities are more important than others. A heart failure patient is likely to be compromised in gross motor activities, whereas fine motor activities will likely be unaffected. On the other hand, a rheumatoid arthritis patient may have difficulty with fine motor activities, whereas cardiopulmonary function may be less affected. The PROMIS data only allowed for a limited number of tests of this issue. Another question is whether a CAT-based approach, which puts emphasis on the unidimensionality of a construct, may be disadvantageous over a classical sum score which may more easily combine different subcategories of Physical Function.

The PROMIS Assessment Center allows users to choose among three types of instruments: 1) Pick-a-Pro (off-the-shelf static forms), 2) Build-a-Pro (user selects items for static forms), and 3) CATs. Pick-a-Pro forms may show advantages or disadvantages compared to legacy instruments; this empirical question will be informed over time. Build-a-Pro is a new approach and *a priori* validation data will not be available for any particular instrument. The advantage of Build-a-Pro is that, for example, a rheumatologist can pick different items from a cardiologist, but both instruments will be comparable. However, if a researcher picks less appropriate items, treatment effects might be overlooked. CAT provides its own challenges because different items may be applied before and after successful treatment or to treatment and control groups. In addition, while in theory an IRT score can be achieved from any combination of items, in practice items from one sub-domain may be more relevant and responsive in a particular disease. Some CAT software provides content balancing to force the CAT to apply the most informative items from predefined sub-domains to estimate one common score. Real-time evaluation of response consistency could evaluate the adequacy of such a balanced score [39]. If an individual's response pattern differs from the model, the CAT would omit a total score and automatically report scores for each sub-domain instead (Figure 2). Ultimately, we think this can be an important advantage of CATs.

4.2 Empirical Issues

PROMIS item banks studied to date show relatively high item discrimination parameters compared to previous studies [7;28]. One possible explanation is that thorough item development resulted in improved items. However, data skewness may have contributed to this as well because the majority of participants did not have serious health issues. Thus, the item bank may seem more consistent than it would be in more disabled samples. Also, PROMIS chose to simplify the assessment by keying all items in one direction, which may have led to response set, with higher inter-item correlations and higher item discrimination parameters.

An issue inherent to item bank development is the balance among different aspects of the latent trait. The content of the majority of items in a bank will have a prominent impact on construct definition. Because upper body items are only one-quarter of the PROMIS Physical Function bank, they carry less item information and would be picked less often by

a CAT, if decisions are based entirely on psychometric criteria. However, validity of an instrument depends on content as well as measurement properties. Upper body items provide particularly good information in the lowest range of function. If we had separated the Physical Function bank into upper and non-upper banks, we would have narrowed the measurement range for both banks. Further there is a clear practical advantage to having one Physical Function score. The use of multidimensional IRT models may be a promising answer to this issue in the future [4], and researchers currently have the option to analyze upper or lower body items separately [40].

4.3 Limitations

An important advantage of generic PRO tools is comparability of results between different diseases. The Wave 1 data only allowed for partial evaluation of the impact of different diseases on item parameter estimates. Thus, additional research is needed to support the assumption that the item banks can be used across patient groups.

While DIF analyses showed that almost all items could be used across disease groups, DIF for a variety of diseases could not be tested for many items. In addition, diseases with known impact on Physical Function, such as back pain, were not included. Disease also was treated as a dichotomous value, but disease severity is in many cases more important in evaluating DIF.

4.4 Perspective

The PROMIS initiative is the largest effort worldwide to improve PRO measures and facilitate their use in clinical research and practice. While initial empirical results are promising, a number of important issues need to be addressed, and many opportunities and challenges of CAT are just being discovered. Most of the issues discussed above also are relevant for instruments developed using classical test theory, but the current interest in IRT-based tools allows for addressing them with rigor. We hope that the current PROMIS item banks can serve in this respect as a starting point for the standardization of PRO measures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Patient-Reported Outcomes Measurement Information System (PROMIS; www.nihpromis.org) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Northwestern University PI: David Cella, PhD, U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project are Deborah Ader, Ph.D., Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, and Susana Serrate-Sztein, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review See the web site at www.nihpromis.org for additional information on the PROMIS cooperative group.

We wish to acknowledge the contribution of Janine Devine (maiden name Janine Becker) as well as Ethan Aaronson to the data analysis, and would like to thank Karen Cook and Ron Hays for their thoughtful comments.

Reference List

1. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care. 2007; 45:S3–S11. [PubMed: 17443116]
2. Cella D, Riley W, Stone A, 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; 63:1179–1194. [PubMed: 20685078]
3. van der Linden, WJ.; Hambleton, RK. Handbook of Modern Item Response Theory. Berlin: Springer; 1997.
4. Wainer, H.; Dorans, NJ.; Eignor, D., et al. Computerized Adaptive Testing: A primer. 2. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
5. Ware JE Jr, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med Care. 2000; 38:II73–II82. [PubMed: 10982092]
6. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol. 2008; 61:17–33. [PubMed: 18083459]
7. Bjorner JB, Kosinski M, Ware JE Jr. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT™). Qual Life Res. 2003; 12:913–933. [PubMed: 14651412]
8. Granger CV, Hamilton BB, Linacre JM, Heinemann AW, Wright BD. Performance profiles of the functional independence measure. Am J Phys Med Rehabil. 1993; 72:84–89. [PubMed: 8476548]
9. Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol. 1994; 47:671–684. [PubMed: 7722580]
10. Fisher WP Jr, Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HSI Physical Functioning Scales. J Outcome Meas. 1997; 1:329–362. [PubMed: 9661727]
11. Raczek AE, Ware JE, Bjorner JB, et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol. 1998; 51:1203–1214. [PubMed: 9817138]
12. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum. 2004; 50:3296–3305. [PubMed: 15476213]
13. Haley SM, Coster WJ, Andres PL, et al. Activity outcome measurement for postacute care. Med Care. 2004; 42:I49–I61. [PubMed: 14707755]
14. Ware JE Jr, Gandek B, Sinclair SJ, Bjorner JB. Item response thoery and computerized adaptive testing: implications for outcomes measurement in rehabilitation. Rehabilitation Psychology. 2005; 50:71–78.
15. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007; 45:S22–S31. [PubMed: 17443115]
16. Dewalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care. 2007; 45:S12–S21. [PubMed: 17443114]
17. Bruce B, Fries JF, Ambrosini D, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther. 2009; 11:R191. [PubMed: 20015354]
18. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980; 23:137–145. [PubMed: 7362664]
19. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992; 30:473–483. [PubMed: 1593914]

20. Liu H, Cella D, Gershon R, et al. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. J Clin Epidemiol. 2010; 63:1169–1178. [PubMed: 20688473]
21. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res. 2009; 18:873–880. [PubMed: 19543809]
22. Rothrock NE, Hays RD, Spritzer K, Yount SE, Riley W, Cella D. Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol. 2010; 63:1195–1204. [PubMed: 20688471]
23. Masters GN, Wright BD. The essential process in a family of measurement models. Psychometrika. 1984; 49:529–544.
24. Muraki, E. A Generalized Partial Credit Model. In: Linden, WJ.; Hambleton, RK., editors. Handbook of Modern Item Response Theory. Berlin: Springer; 1997. p. 153-164.
25. Muthén, LK.; Muthén, BO. Mplus. The Comprehensive Modeling Program for Applied Researchers. User's Guide. Los Angeles: Muthén & Muthén; 1998.
26. Nunnally, J. Psychometric Theory. 2. New York: MacGraw-Hill; 1978.
27. Reckase M. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics. 1979; 4:207–230.
28. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Qual Life Res. 2005; 14:2277–2291. [PubMed: 16328907]
29. Stout W. Psychometrics: From practice to theory and back: 15 years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment. Psychometrika. 2002; 67:485–518.
30. Nagelkerke NJD. Miscellanea. A note on a general definition of the coefficient of determination. Biometrika. 1991; 78:691–692.
31. Ramsay, JO. TestGraf. A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data. Montreal: McGill University; 1995.
32. Orlando M, Thissen D. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. Appl Psych Measur. 2000; 24:50–64.
33. Bjorner, J.; Smith, K.; Stone, C.; Sun, X. IRTFIT: a macro for item fit and local dependence tests and IRT models. Lincoln, RI: QualityMetric; 2007.
34. Muraki E. Information functions of the generalized partial credit model. Appl Psych Measur. 1993; 17:351–363.
35. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res. 2009; 18:447–460. [PubMed: 19294529]
36. Broderick J, Schneider S, Junghaenel D, Schwartz J, Stone A. Validity and relaibility of patient-reported outcomes measurement information system (PROMIS) instruments in osteoarthritis. Arthritis Care Res. 2013; 65:1625–1633.
37. Martin M, Kosinski M, Bjorner JB, Ware JE Jr, Maclean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res. 2007; 16:647–660. [PubMed: 17334829]
38. Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. Med Care. 2007; 45:S32–S38. [PubMed: 17443117]
39. Emons WH, Sijtsma K, Meijer RR. Global, local, and graphical person-fit analysis using person-response functions. Psychol Methods. 2005; 10:101–119. [PubMed: 15810871]
40. Hays R, Spritzer K, Amtmann D, Lai J, Dewitt E, Rothrock N. Upper-extremity and mobility subdomains from the Patient-Reported Outcomes Measurement Information System (PROMIS) adult physical functioning bank. Arch Phys Med Rehabil. 2013; 94:2291–2296. [PubMed: 23751290]

What is new

The paper describes the development of a comprehensive physical function item bank. This item bank can improve the measurement of physical function by standardizing the metric and enabling short and precise CAT assessment through readily available software, thus facilitating the use of patient self-assessment in clinical practice and research.

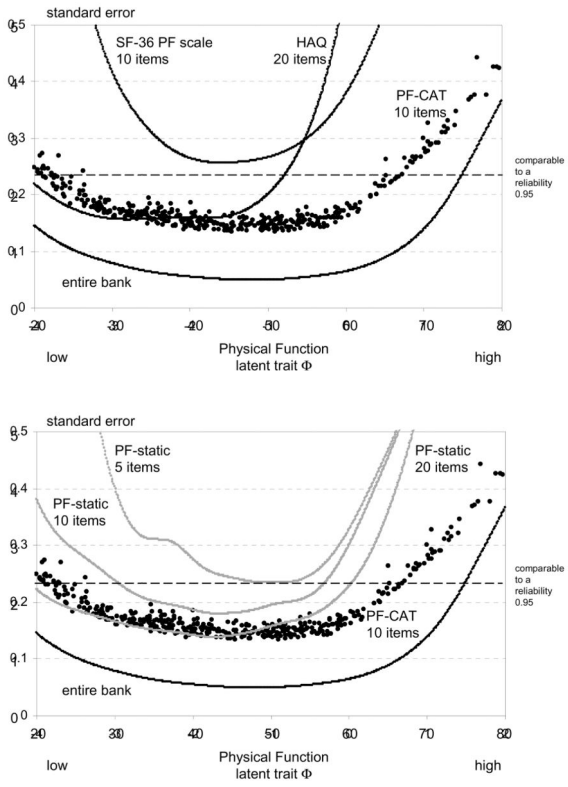

Figure 1. Measurement precision and range of two established static tools, PROMIS static forms, a simulated PROMIS PF-CAT with 10 items, and the entire item bank

The Y-axis shows the standard error of measurement, the X-axis the Θ-value. The graph shows the precision of the test which can be expected at a particular level of Physical Function (latent trait, Θ) based on test information (static forms) or simulation studies (CAT)

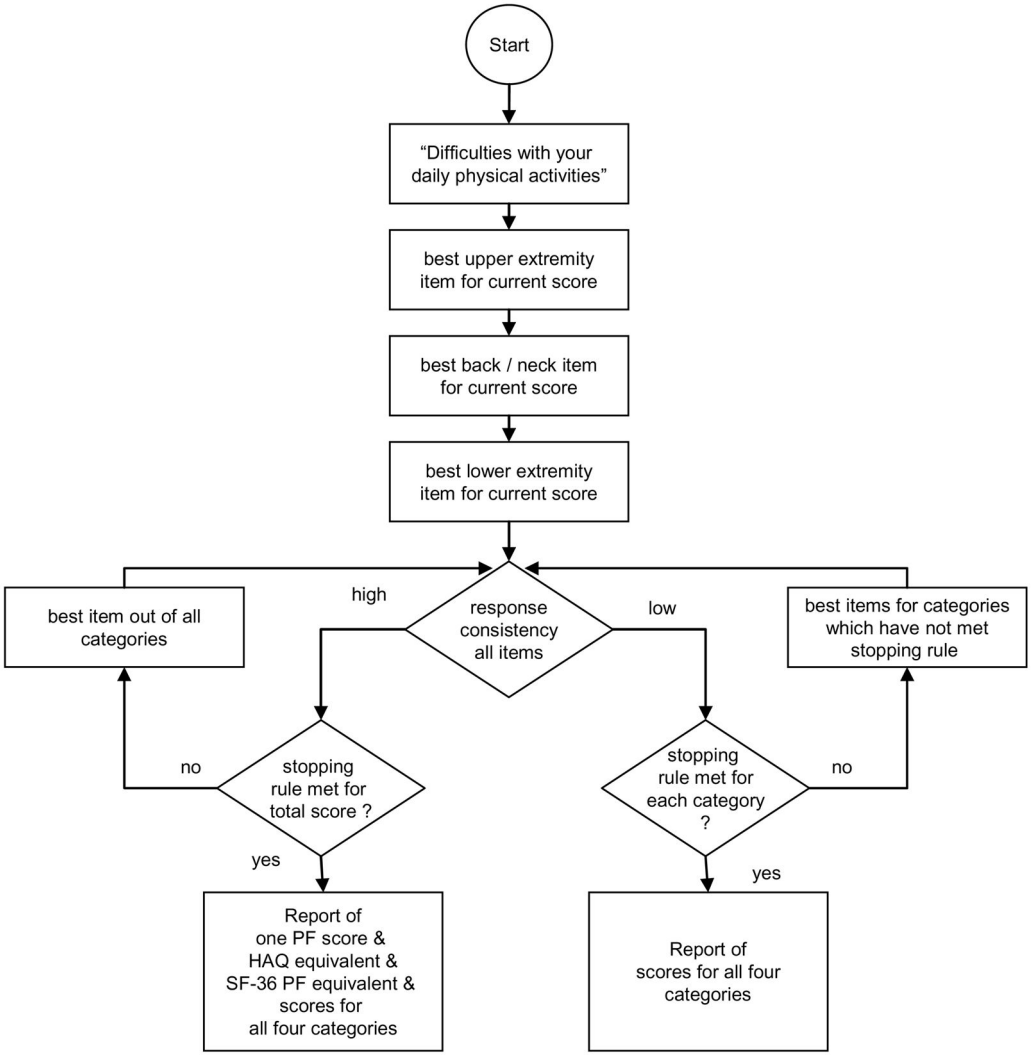

Figure 2.
Proposed CAT Algorithm

Table 1

Item Description and Statistics for the PROMIS Physical Function Item Bank

(20 sample items are shown sorted by Imax at Θ, parameters for all estimated PROMIS items are accessible via the PROMIS website)

	item label and stem [1,2]	a priori dom.	static Form[3]	skew-ness	corr. to global item[4]	CFA[5]	DIF[6]	slope[7]	thresholds[7]				I_{max}[8] at Θ	CAT util. % [9]
									1	2	3	4		
Are you able to …														
A51	sit on the edge of a bed?	Cent.	20	−5.07	0.68	0.88		3.22	13.9	18.7	24.4	29.8	2.25 14.6	0.0 8.6
B25	push open a door after turning the knob?	Upper	20	−5.46	0.75	0.93		3.34	15.9	20.9	26.8	33.3	1.97 15.8	0.0 0.0
C46	transfer from a bed to a chair and back?	Lower	20	−3.63	0.75	0.89		3.61	15.2	20.7	26.7	34.4	1.31 14.6	0.4 24.0
B26	shampoo your hair?	Up	20 10	−6.00	0.74	0.88		3.52	18.7	21.5	27.2	32.7	2.82 17.8	0.2 15.7
C45	get on and off the toilet?	Lower	20 10	−3.95	0.74	0.86		3.18	17.7	21.3	27.4	35.2	2.90 16.5	0.1 2.9
A56	get in and out of a car?	Cent.	20	−2.19	0.75	0.88		3.24	13.6	22.5	30.0	38.9	2.33 20.4	0.0 5.6
A16	dress yourself. including tying shoelaces and doing buttons?	Upper	20 10	−3.36	0.71	0.92	0.04 sex	3.31	18.1	24.4	30.7	37.5	2.70 20.7	0.0 0.0
A11	do chores such as vacuuming or yard work?	IADL	20 10	−1.67	0.81	0.93		4.72	30.1	34.7	39.1	45.8	5.58 32.9	45.7 75.0
B22	hold a plate full of food?	Upper	20	−4.86	0.76	0.91	0.03 edu	3.31	15.7	22.6	27.4	34.3	2.41 22.0	0.0 0.0
B15	change the bulb in a table lamp?	Upper	20	−5.40	0.81	0.94		3.47	26.2	31.9			2.33 22.1	0.0 0.0
A38	dry your back with a towel?	Cent.	20	−3.09	0.74	0.88		2.79	19.4	24.6	30.3	35.9	2.01 22.3	0.0 0.0
B19	squeeze a new tube of toothpaste?	Upper	20	−6.50	0.67	0.93		3.22	20.0	25.5	30.3		2.49 23.2	0.0 0.0
A55	wash and dry your body?	IADL	20 10	−3.16	0.78	0.91		3.52	15.9	22.9	28.8	34.7	3.04 26.2	0.2 12.1
B24	run a short distance. such as to catch a bus?	Lower	20	−1.19	0.80	0.92		3.90	36.2	39.2	42.8	49.0	3.74 38.0	0.6 0.1
Does your health now limit you in …														

	item label and stem [1,2]	a priori dom.	static Form[3]	skew-ness	corr. to global item[4]	CFA[5]	DIF[6]	slope[7]	thresholds[7]				I_{max}[8] at Θ	CAT util. % [9]
									1	2	3	4		
A05	lifting or carrying groceries?	IADL	20 10 5	−1.58	0.85	0.93		3.99	26.0	33.5	39.6	44.9	4.01 38.8	0.0 0.0
C37	climbing one flight of stairs?	Lower	20 10 5	−1.38	0.87	0.94		4.26	26.6	33.4	39.4	44.2	4.71 39.0	0.0 1.3
C36	walking more than a mile?	Lower	20 10 5	−0.77	0.83	0.94		4.23	35.5	39.8	44.1	47.9	4.94 43.5	0.0 0.0
A03	bending. kneeling. or stooping?	Cent.	20 10 5	−0.55	0.78	0.89		2.81	26.7	37.0	44.3	50.5	2.13 44.4	0.0 0.0
C12	doing two hours of physical labor?	IADL	20	−0.71	0.77	0.91		4.49	35.8	40.9	46.3	50.9	5.2 46.1	59.8 31.7
A01	doing vigorous activities. such as running, lifting heavy objects, participating in strenuous sports?	IADL	20 10 5	0.11	0.76	0.94	0.03 age 0.04 dis	2.99	38.2	45.0	51.7	56.5	2.68 52.2	39.3 0.6

[1] Sample size per item ranges from n=2,220 to 2,926.

[2] Complete item text and responses can be found at the PROMIS website http://www.assessmentcenter.net. Response scale for Items that start with “Does your health now limit you…” is: Cannot do, Quite a lot, Somewhat, Very little, Not at all. Response scale for Items that start with “Are you able…” is: Unable to do, With much difficulty, With some difficulty, With a little difficulty, Without any difficulty.

[3] Item included in the 20-item, 10-item and/or 5-item static form.

[4] Correlation with global Physical Function item (labeled as ‘Global’).

[5] Factor loading in one-factor confirmatory factor analysis.

[6] DIF column shows p value for DIF tests where $\Delta R^2 > 0.03$, and indicate type of test: Age (65+ versus <65), Gender (male versus female), Education (high school education or less versus more than high school education), Dis: Disease (musculoskeletal, cardiopulmonary, mental, no condition).

[7] The slope parameter is also called discrimination parameter. Higher slope parameters indicate better discrimination, which makes the item more valuable, i.e. ‘informative’, for score estimation. Thresholds define the range at which a particular response option is most likely to be endorsed. Typically items have one fewer threshold than response options. If item responses needed to be collapsed to ensure a stable IRT estimate for those items, fewer thresholds are established. The mean threshold illustrates the position of an item on the metric, or its ‘item difficulty’ in traditional terms.

[8] I_{max} at Θ: Maximum of the information function (upper number in each cell) at a particular Θ (lower number in each cell).

[9] CAT Util.: The upper number shows the likelihood that the item was used in a sample with distribution characteristics of the general population (mean=50. SD=10), the second that the item was used in a sample with distribution characteristics of a clinical population (mean=30, SD=10). The CAT start item was the Global item (To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair? Not at all, A little, Moderately, Mostly, Completely).

Table 2

Scores in Relation to Number of Chronic Conditions and General Health Rating

			Item Bank		Simulated 10 Item CAT		20[1] Item Static Form		10 Item Static Form		5 Item Static Form		SF-36 PF-10 Scale		HAQ-DI	
	Group	**N**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
General Health	Poor	27	35.1	9.4	35.5	10.7	27.5	15.4	27.8	12.6	31.5	9.3	26.4	12.1	76.2	18.7
	Fair	99	42.8	8.8	43.2	8.9	41.5	13.6	41.2	12.5	41.3	10.9	38.9	11.9	56.2	14.2
	Good	301	48.7	8.8	48.7	8.7	49.5	8.2	49.0	8.7	48.5	9.3	45.5	10.6	50.0	8.0
	Very good	278	54.4	7.6	54.6	7.8	53.6	5.8	53.8	6.0	54.0	6.8	51.9	6.9	47.1	5.8
	Excellent	114	58.0	8.0	58.2	8.2	54.8	4.2	55.3	5.1	55.7	6.2	53.6	6.5	46.4	3.1
		F(rv)[2]	86.2 (1.00)		82.6 (0.96)		91.3 (1.06)		91.5 (1.06)		84.1 (0.98)		76.9 (0.89)		77.6 (0.90)	
Age	<=20	32	56.4	7.6	56.9	7.3	54.6	8.7	55.2	8.5	56.6	6.8	53.2	6.7	47.6	9.3
	20–39	241	55.4	8.4	55.8	8.5	53.9	6.1	54.2	6.4	54.7	6.9	51.5	8.1	47.4	5.9
	40–59	271	50.8	10.2	51.1	10.1	49.8	11.1	49.8	10.8	50.0	10.6	47.7	11.6	50.7	11.6
	>=60	274	46.0	8.8	45.9	8.9	46.4	10.2	46.0	10.2	45.2	9.6	43.4	11.2	51.7	10.6
		F(rv)[2]	50.4 (1.00)		55.8 (1.11)		27.0 (0.54)		32.0 (0.63)		47.2 (0.94)		24.9 (0.49)		8.2 (0.16)	
Chronic Conditions	None	356	55.6	8.1	55.8	8.3	53.9	6.0	54.2	6.5	54.6	7.2	52.2	7.5	47.2	5.9
	One	179	51.7	8.2	51.9	8.5	51.9	6.7	52.1	6.6	51.6	7.9	49.5	8.2	47.7	5.4
	Two or more	284	44.0	9.1	44.3	9.1	43.6	12.6	43.1	11.9	43.1	10.5	39.9	12.4	55.4	13.9
		F(rv)[2]	147.3 (1.00)		139.8 (0.95)		102.9 (0.70)		122.1 (0.83)		137.3 (0.93)		118.7 (0.81)		60.2 (0.41)	

[1] For better illustration of T-score transformation of sum scores for the static forms: 20 items: $y_i = (((x_i - 4.4367)/.50840) *10) + 50$; 10 items: $y_i = (((x_i - 4.4600)/.66385) *10) + 50$; 5 items: $y_i = (((x_i - 4.0155)/1.02460) *10) + 50$, and the HAQ-DI $y_i = (((x_i - 2.5354)/5.71807) * 10) + 50$, SF-36 norm-based scoring as recommended by the test authors

[2] F: ANOVA F-value and rv: relative validity (F-value derived by other instrument/F-value derived by IRT bank score)