Published in final edited form as:
Arch Phys Med Rehabil. 2011 October ; 92(10 0): S20–S27. doi:10.1016/j.apmr.2010.08.033.

How Item Banks and Their Application Can Influence Measurement Practice in Rehabilitation Medicine: A PROMIS Fatigue Item Bank Example

Jin-Shei Lai, Ph.D., OTR/L[*], **David Cella, Ph.D.**[*], **Seung Choi, Ph.D.**[*], **Doerte U. Junghaenel, Ph.D.**[‡], **Christopher Christodoulou, Ph.D.**[‡], **Richard Gershon, Ph.D.**[*], and **Arthur Stone, Ph.D.**[‡]

[*]Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, Illinois

[‡]Department of Psychiatry and Behavioral Sciences Stony Brook University, Stony Brook, New York

Abstract

Objective—To illustrate how measurement practices can be advanced using as an example the fatigue item bank (FIB) and its applications (short-forms and computerized adaptive test) that were developed via the NIH Patient Reported Outcomes Measurement Information System (PROMIS) Cooperative Group.

Design—Psychometric analysis of data collected by an internet survey company using Item Response Theory (IRT) related techniques.

Setting—A United States general population representative sample collected via internet.

Participants—803 respondents used for dimensionality evaluation of the PROMIS FIB and 14,931 respondents used for item calibrations

Interventions—Not applicable.

Main Outcome Measures—112 fatigue items developed by the PROMIS fatigue domain working group, 13-item Functional Assessment of Chronic Illness Therapy-Fatigue, and 4-item SF-36 Vitality scale.

Results—The PROMIS FIB version 1 which consists of 95 items demonstrated acceptable psychometric properties. Computerized Adaptive Testing (CAT) showed consistently better precision than short-forms. However, all three short-forms showed good precision for the majority of participants, in that more than 95% of sample could be precisely measured with a reliability greater than 0.9.

Conclusions—Measurement practice can be advanced by using a psychometrically sound measurement tool and its applications. This example shows that CAT and short-forms derived

Corresponding Author: Jin-Shei Lai, Ph.D, OTR/L, Research Associate Professor, Medical Social Sciences and Pediatrics, Northwestern University Feinberg School of Medicine, 710 N Lake Shore Drive, #724, Chicago, IL 60611, TEL: 312-503-3370, FAX: 312-503-6743.

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

from the PROMIS FIB can reliably estimate fatigue reported by the US general population. Evaluation in clinical populations is warranted before the item bank can be used for clinical trials.

Keywords

PROMIS; fatigue; CAT; short-form

Measuring patient health-related quality of life and symptoms in a brief-yet-precise manner has been a challenge. Given the rapid growth of the computer technology and the advances in modern test theory such as item response theory (IRT), the patient-reported outcomes (PRO) assessment field has matured to a point where patient assessment is rapidly becoming computerized. Computer-based testing allows for more frequent assessments and immediate feedback with minimal burden on patients and providers.[1] With available IRT-calibrated item banks including large numbers of questions, static short forms and computer adaptive testing (CAT) of varying lengths can be constructed using various strategies to efficiently estimate a person's score on a unidimensional measure.

More than a collection of items, an *item bank* is comprised of items calibrated by the item response theory (IRT) models.[2–5] The items in a bank are concrete manifestations of positions along that continuum that represent differing levels of that trait. A psychometrically-sound item bank can provide a basis for designing the best set of questions for any particular application. An IRT-calibrated item bank makes it possible to compare the trait levels of two patients who respond to different sets of questions in the bank. A significant advantage of an item bank is that it provides the foundation for the development of dynamic CAT platforms and static fixed length short-forms.[6, 7] Fixed-length short-forms in which a subset of bank items can be selected from across the trait spectrum to produce a static instrument can be used when access to computers is limited. The scores produced by any of the instruments created from the calibrated bank are calibrated on the same continuum and are comparable regardless of the specific questions asked of a given individual or group of respondents.[8, 9] Computerized adaptive testing is a dynamic process of test administration in which items are selected on the basis of the patients' responses to previously administered items.[10] This process utilizes a computerized algorithm to custom select the most informative items from the item bank which is targeted on the estimated person level (e.g., fatigue), where estimated person level is based upon the patient's previous responses at each point in the test. The CAT is further administered under specific test specifications, such as content coverage and test length. For example, it allows fine-grained assessment of those with both low and high levels of the construct by presenting questions appropriate for each person (many low-difficulty questions for the former person and many high difficulty questions for the latter). In this paper using fatigue as an example, we demonstrated how applications from a psychometrically sound item bank can enhance the rehabilitation practice.

Fatigue is a common complaint for people with chronic illness seen in rehabilitation settings and a potential cause of disability in may disease processes such as cerebral palsy,[11] cardiopulmonary disease,[12] rheumatology,[13] stroke,[14] and multiple sclerosis.[15] Using cancer as an example, depending on the criterion and the assessment tools being used, the prevalence rates ranged from 18% to 96%.[16, 17] Not only for people with chronic illness, approximately 20% of men and 30% of women in the general population complain of frequent tiredness.[18] As a symptom, fatigue is defined as a subjective sensation of weakness, lack of energy or tiredness.[19] As a syndrome, it has been defined as an overwhelming, sustained sense of exhaustion and decreased capacity for physical and mental work.[20]

Fatigue can be distinguished as primarily physiological (e.g., muscle strength, exercise tolerance, or maximal oxygen capacity after exercise) or self-report (i.e., patients' perceptions of fatigue and its consequences). There have been an on-going debate whether fatigue should be considered as uni-dimensional or multi-dimensional and various scales have been developed accordingly, such as the Functional Assessment of Chronic Illness Therapy - Fatigue (FACIT - F),[21, 22] the Brief Fatigue Inventory,[23] the Piper Fatigue Scale,[24] the Multidimensional Fatigue Inventory,[25] and the Fatigue Symptom Inventory.[26] Lai and her colleagues[6, 27] evaluated dimensionality of fatigue using various approaches. They concluded fatigue is sufficiently unidimensional from a measurement's perspective and it is reasonable to report fatigue by using a single score. A brief-yet-precise fatigue assessment that is easily implemented in clinics is critical to facilitate early identification of fatigue experienced by patients. Applications from a comprehensive fatigue item bank, such as computerized adaptive tests (CAT) and short-forms, are the best candidates to fulfill this goal.

The Patient Reported Outcomes Measurement and Information System (PROMIS, www.nihpromis.org) is a National Institute of Health (NIH) Roadmap initiative to develop item banks to measure patient-reported symptoms and other aspects of health-related quality of life across various condition and disease populations, including patients commonly seen in rehabilitation clinics. To date, PROMIS has developed numerous items banks including physical function, emotional distress, pain, social function and fatigue. This paper illustrates the development of the PROMIS fatigue item bank version 1 (FIB), the resulting short-forms and CAT, and their potential contributions to the rehabilitation practice.

METHOD

Sample

Data obtained from PROMIS Wave 1 testing were used for this study. The sampling plan is documented in Cella et al (in the same issue). The full bank sample (i.e., participants who were assigned to complete all items included in the fatigue item pool) was used to determine dimensionality of fatigue, while both full bank and block data (i.e., participants who completed 7 items measuring fatigue experience, 7 fatigue impact and also 7 items from each of all other 12 domains included in the PROMIS Wave 1 testing) were used for item parameter estimation. These 12 domains are; anxiety, depression, alcohol abuse, anger, physical function, fatigue experience, fatigue impact, social health/role performance, social health/ role satisfaction, pain interference, pain quality, and pain behavior. In the full bank analyses, participants were included in the analysis if 1) they responded to 50% or more of the items; 2) they did not have repetitive strings of 10 or more identical responses; and 3) their response time was greater than 1 second per item.

Fatigue Item Pool Generation

The initial PROMIS fatigue item pool consisted of 112 items tapping two conceptual areas: individual's fatigue "experience" and impact of fatigue on an individual's daily living (i.e., "impact"). All items were drafted from literature review, patient focus groups and individual cognitive interviews.[28] The "experience" bank contained 54 items measuring intensity, frequency or duration of fatigue; while "impact" bank contained 58 items measuring impact on physical function, emotional function or social function. Five-point rating scales were used to measure intensity (Not at all/A little bit/Somewhat/Quite a bit/Very much), frequency (Never/Rarely/Sometimes/Often/Always), or duration of fatigue (None/1 day/2–3 days/4–5 days/6–7 days). Additionally, the 13-item Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F)[21] and 4-item SF-36 Vitality Scale[29] were included in the PROMIS fatigue testing and are referred to as "legacy" measures in the remaining of this paper. “Legacy measures” in the PROMIS are referred to those widely used fixed measures in the field. Selection of the legacy measures is determined by each domain working group and the PROMIS Steering Committee. We selected FACIT-F and SF-36 Vitality Scale due to their solid psychometric properties. Legacy measures were included in the calibrated item banks to facilitate the cross-walk between PROMIS item banks and existing database established by other studies.

Analysis

As there is no common single indicator agreed upon by psychometricians to evaluate unidimensionality of items, several approaches were implemented to flag candidates for removal. The final decision was determined by the PROMIS fatigue working group. Loevinger’s H as implemented in the Mokken Scale Procedures[30] was used to investigate scalability at both item and scale levels. An item (or scale) would be considered to have weak scalability when $0.3 \leq H < 0.4$, medium when $0.4 \leq H < 0.5$, and strong scalability when $0.5 \leq H$.[31] Unidimensionality of items was evaluated by bi-factor analysis, a family member of confirmatory factor analysis (CFA). Common criteria for fit statistics include root mean squared error of approximation (RMSEA) <0.08 for adequate fit and <0.05 for good fit; Tucker-Lewis index over 0.90 or 0.95; and comparative fit index (CFI) over 0.90 or 0.95.[32] The challenge of dimensionality assessment is to develop approaches that assess whether a scale has a strong enough general factor to be considered *sufficiently unidimensional*, because no complex item set will ever perfectly meet traditionally strictly defined unidimensionality assumptions.[33] What we are really interested in assessing is whether the trait level estimates are predominantly influenced by a general factor. Cook et al[34], based on a series of simulations with various data distributions and numbers of items, concluded that traditionally used fit values are sensitive to influences other than dimensionality of the data and recommend the use of bi-factor analysis as an adequate and informative approach for developing an item bank. Bi-factor analysis includes two classes of factors: a general factor, defined by loadings from all of the items in the scale, and local factors, defined by loadings from pre-specified groups of items related to that sub-domain.[27, 33, 35, 36] Items are considered sufficiently unidimensional when standardized loadings are > 0.3 for all the items on the general factor. Similarly, if the loadings of all the items on a local factor are salient, this would indicate that the local factor is well defined even in the presence of the general factor, and it is more appropriate to report scores of local factors separately.[27, 33, 37] The hypothesized model shown in Figure 1 was used for bi-factor analyses. Further, item fit was evaluated by using S-χ^2 and S-G^2 developed by Orlando and Thissen,[38] which compare the predicted and observed response frequencies for each level of the scale sum score. Items with p<.001 are considered poorly fitting the data and are candidates for removal. Differential item functioning (DIF) by gender and age were evaluated by using ordinal logistic regression, OLR,[39] and a statistical test developed by Mantel[40] as described in Zwick and Thayer.[41] The Mantel chi-square DIF statistic is an extension of the Mantel-Haenszel (MH) test of conditional association.[42] Like the MH test, the Mantel DIF test uses a stratification variable, which is often defined as the sum of item scores. The null hypothesis states that when members of the reference and focal groups are matched on stratification variation (i.e., test score), they tend to show the same item scores. For both OLR and MH, items with p< 0.01 were considered demonstrating DIF. Items that demonstrated DIF on both gender and age at p <0.01 were not included in the final version of the PROMIS fatigue item bank. Exclusion/inclusion of items that showed DIF on one of the DIF approaches was determined by the PROMIS fatigue working group to ensure the clinical relevance of produced fatigue item bank. Finally, item parameters were estimated using the Graded Response Model[43] as implemented in MULTILOG, which takes both slopes (discrimination) and threshold parameters (difficulty) into account as we were interested in both item parameters, and did not intend to exclude discrimination function from the final information function estimation. All psychometric approaches chosen by the PROMIS psychometric team are described elsewhere[44] and are not re-introduced here with the consideration of the length of the manuscript.

We then compared the precision levels along the fatigue continuum, defined by item parameters of the PROMIS FIB, among three short-forms and CAT by using the error functions, converted by scale information function. Details of short-form construction and CAT platform are described in the next sections. An information function (IF) is the reciprocal of the standard error function. It indicates the maximum accuracy with which a patient's fatigue level is estimated at different points along the fatigue continuum, which varies depending on the location on the fatigue continuum.[45–47] High information functions correspond to high precision of the fatigue estimates (i.e., low error) and vice versa. We anticipated that these three short-forms would demonstrate different levels of precisions along the continuum: one more precise for people with moderate fatigue (middle of the fatigue continuum), another more precise for people with severe fatigue, and the other for people with mild fatigue. Yet, CAT was expected to show better precision than all short-forms.

Construction of a Fatigue Short-Form

There are many methods available to construct short-forms. For this paper, three short-forms were constructed for illustration purposes. The first short-form was content-oriented, in which seven items were selected by multidisciplinary panels of clinical experts (including physicians, nurses, pharmacists, and psychologists). Details of this effort are reported in Garcia et al.[48] The other two short-forms were created purely from a measurement perspective, regardless of content, by using item parameter threshold values obtained from IRT estimation. The "high-end" short-form consists of seven items, which had the highest calibrations on the fatigue continuum (i.e., items reflecting the most severe fatigue) and were not part of the content-oriented short form. This short-form was constructed for use with people with more severe fatigue. The "low-end" short-form was composed of seven items with the lowest calibrations on the fatigue continuum (i.e., items reflecting less fatigue) in order to capture people with mild fatigue. For comparison purposes, all three short-forms consist of seven items.

Computerized Adaptive Test Platform

We used the first generation PROMIS CAT engine, Firestar,[49] in this study. More than one method is available to choose the first item. In this study, we chose to start CAT with the item with the maximum information function at the distribution mean (theta=0). Participants' scores are re-estimated according to his/her endorsement. Items with the maximum information function at the re-estimated scores are chosen as the subsequent item to be administered by the CAT engine. This estimation process continues until the standard error is < 0.3 or the number of items administered is > 20, whichever comes first. The engine employs an Expected A Posteriori (EAP) theta estimator and the maximum posterior weighted information (MPWI) item selection criterion. The MPWI[50] selects items based on the information function weighted by the posterior distribution of trait values. Choi and Swartz[51] demonstrated that MPWI in conjunction with EAP provides excellent measurement precision for polytomous item CAT, marginally superior to the traditional maximum information criterion and comparable to other computationally intensive Bayesian selection criteria. The CAT can be administered on-line or via a stand-alone computer. Thus, patients can complete CAT testing in clinical settings or at home or any place as long as internet access is available.

RESULTS

Sample

Consistent with all other item banks developed via the PROMIS initiative, full-bank data were used to evaluate unidimensionality of items while full-bank plus block data were used for final item calibrations.

Full bank data (used for evaluating dimensionality)—The average age of the 803 participants was 51.8 (SD=17.8; range: 18–89). 55% of the sample was female and 45% was male; 11% was of Hispanic origin; 81% was white, followed by 10% African American and 9% multiple races. In terms of education, 20% high school or lower, 44% some college, 18% college, and 18% advanced degree. Thirty-four percent of participants reported having a diagnosis of hypertension, 22% arthritis or rheumatism, 20% depression, 15% anxiety, and 14% asthma, 14% OA or degenerative arthritis, and 14% migraines or severe headache.

Calibration Sample (Full bank and block data combined used for item calibration)—The average age of the 14,931 participants was 54.1 (SD=16.4; range: 18–100). 52% of the sample was female and 48% was male; 8% was Hispanic origin; 83% was white, followed by 8% African American and 8% multiple races. In terms of education, 18% high school and under, 38% some college, 24% college, and 20% advanced degree, 12% with family household income < $20,000, 33% between $20,000 and $49,000, 36% between $50,000 and $99,000, and 19% > $100,000. Forty-three percent reported having hypertension, 28% arthritis or rheumatism, 28% depression, 20% cancer, 18% migraines or severe headache, 18% anxiety, 18% OA or degenerative arthritis, 17% asthma, 15% sleep disorder, 15% diabetes, 14% COPD, bronchitis, emphysema, and 12% angina. All other diseases were reported by less than 10% of the sample.

Analysis Results

Multiple steps were taken to build the PROMIS fatigue item bank version 1. Figure 2 summarizes our decision making processes with associated analysis results. Decisions made at each step were based on both psychometric analysis results as well as clinical relevance determined by the PROMIS fatigue domain group. “Impact” and “experience” items were initially analyzed separately and then analyzed together to examine whether fatigue could be reported by using a single score. Items from legacy scales (i.e., 13 from the FACIT-F and 4 from SF36/Vitality) were included in analysis of fatigue experience to enable the cross-walk between the final PROMIS bank and other disease groups.

For “fatigue impact”, Spearman’s rhos were greater than 0.3 for all except those related to two fatigue impact items. The same two items also showed low item-scale correlations and low H in Mokken Scale Procedures analysis and, therefore, were removed from the item pool. Strong scalability (H=0.7) was found when these two items were removed. Five item-pairs showed residual correlations greater than 0.2, suggesting potential local dependency within these pairs; one item of each local of the dependent item-pairs was therefore set aside. No item was rejected using the fit statistics ($S\text{-}\chi^2$ and S-G[2,38] p >0.01), supporting the unidimensionality of these items. For “fatigue experience”, seven items showed Spearman’s rho less than 0.3. Results of the Mokken Scale Procedures showed all items had acceptable H and the scale coefficient H of 0.7 indicated strong scalability among these items. However, 10 items were removed due to the local dependency concern. Two more items were rejected by $S\text{-}\chi^2$ and $S\text{-}G^2$, p <0.01, and therefore were removed from the further analysis.

Sufficient Unidimensionality of Overall Fatigue—Seven impact items and 10 experience items were excluded based on the analyses reported above. Bi-factor analysis results showed that all items had higher loadings on the general factor (i.e., overall fatigue) than on their local factors (i.e., either "experience" or "impact"). Fit indices were: CFI=0.91; TLI=1.00; RMSEA=0.10. We were not surprised to find out the RMSEA is slightly higher than the common RMSEA criterion given the large number of items included in the PROMIS FIB. Considering the factor loadings and the fit indices, we concluded that the sufficient unidimensionality of the PROMIS FIB was confirmed. A high Pearson's correlation (r=0.95) between "impact" and "experience" confirmed that these two concepts could be conceptualized as measuring the same underlying construct and could be scaled together.

A post-hoc content review was conducted to ensure the integrity of the remaining items. The PROMIS fatigue domain working group decided to remove three items from the pool after the re-evaluation of item content. In addition, 7 items were rejected by S-χ^2 and S-G^2, p <0.01 when the remaining items were re-analyzed together by using the calibration sample.

Differential Item Functioning (DIF)

Four items exhibited DIF on both gender and age when the calibration sample was analyzed by using ordinal logistic regression. Results showed no items with gender DIF and eight items with age DIF according to OLR; MH results identified six items with gender DIF and 10 with age DIF. Though no item was identified having DIF on both gender and age by using both OLR and MH, The PROMIS fatigue working group decided to remove four items that showed both gender and age DIFs by using MH and showed age DIF by using OLR after reviewing these items. Finally, three items with potential intellectual property concerns were set aside and not included in the final calibration. As a result, the PROMIS fatigue item bank version 1 consists of 95 items,

Relationship with the Legacy Items

We then compared the non-legacy items in the PROMIS fatigue item bank (v1) to the legacy items (i.e., FACIT-fatigue and the SF-vitality scale). Legacy items were not included in the FIB when the comparisons were conducted. High correlations were found on all comparisons. FACIT-fatigue was significantly correlated with "impact" and "experience", r=0.93 and 0.94, respectively. Similarly, the SF-36/vitality scale had noteworthy correlations with the fatigue "impact" and "experience" scores, r=0.85 and 0.90, respectively.

Comparisons of CAT and Short-Forms

A post-hoc CAT simulation was conducted using the full-bank data without missing responses. The length of CAT was kept the same as the length of the short forms. Therefore, the CAT stopped after administering the 7th item regardless of the level of error associated with the fatigue estimate. The first item was selected to provide the maximum expected information over the prior distribution (a normal distribution). The subsequent items were selected targeting the posterior distributions, which converge to progressively narrow ranges as more items are administered. However, we incorporated a random component in item selection to promote more diverse selection of items for the bank. That is, instead of selecting the best item all the time, we selected one at random from a set of seven best candidate items at each stage. Because the Fatigue bank was large, incorporating the random component in item selection had a negligible impact on the measurement precision.

We estimated the error functions of three short-forms and CAT by using Firestar,[49] a CAT simulation program. Results are shown in Figure 3. All IRT-based scaled scores were converted into T-Scores (mean=50 and standard deviation=10), matching the 2000 US

Census data by race, gender and education. As expected, CAT showed consistently better precision than short-forms. This is shown in the graphic by the relatively low (compared to other forms of administration) standard errors throughout a wide range of T-scores (in other words, very good estimates of the construct in those with a wide range of fatigue). Among these three short-forms, better precision was found depending on their designed target areas; the low-end short form was more precise for people with less fatigue (higher SEs in people with high fatigue levels), and high-end for people with more fatigue while content-oriented for people with moderate fatigue. However, all three short-forms showed good precision for the majority of participants. Specifically, the low-end short-form had reliability greater than 0.9 for people with T-scores 32–73, content-oriented for people with T-scores of 40–78, and high-end short-form for people with T-scores of 45 and higher. Ninety-five percent of participants could be estimated with a reliability of 0.9 or higher by all short-forms (T-score=32 at the 5th percentile). A nearly normal distribution was found.in the person fatigue T-scores (Figure 4).

DISCUSSION

Static scales in today's generic and targeted instruments are almost universally too coarse for individual classification and diagnosis. Current trends in measuring health-related outcomes in rehabilitation medicine have been moving towards the use applications of the IRT based item banks, mainly CAT and short-forms. This paper documents the development and psychometric properties of the PROMIS FIB. Using CAT simulation software, we found that the PROMIS FIB based CAT can estimate self-reported fatigue in a very precise manner across the fatigue continuum. Scores obtained from CAT and short-forms were found to have reliability of 0.9 or higher for 95% of participants. CAT has important contributions to clinical settings. CAT results can be readily translated into real-time reports of health-related quality of life for immediate use by the health care providers. The integration of systematic health related quality of life data into regular clinical cancer care can facilitate better symptom management.[52] However, although computer usage has become a daily ritual for many Americans, not all clinical settings have access to computers making investigators not able to take advantages of CAT. In such cases, fixed length short-forms would be ideal alternatives. Multiple short-forms can be developed and their results can be compared as long as items are all derived from a well-calibrated item bank. In this paper, we provide three example short-forms: one content-oriented and the other two measurement-oriented. As expected, although short-forms are not as precise as CAT, they all demonstrated excellent precision along the trait continuum and therefore can serve as reliable alternatives for CAT.

Investigators can create their short-forms to meet specific research or clinical needs. A look-up table can be created and clinicians can easily compare patients' fatigue scores to the US general population as all items are calibrated onto the same continuum. The look-up table for the content-oriented short-form is available by logging-in the PROMIS Assessment Center (http://www.assessmentcenter.net/ac1). Another example look-up table to cross-walk FACIT-Fatigue score to the PROMIS T-scores matrix can be found in Smith, Lai and Cella.[53]

All psychometric information of the PROMIS FIB is publicly available at Assessment Center™ (http://www.assessmentcenter.net/ac1). We have transferred all IRT based scaled scores into T-scores (mean=50, standard deviation, SD = 10). Therefore, clinicians, or even patients, can know patients' fatigue status comparing to the national norm. For example, a patient with a fatigue T-score=71 means his/her reported-fatigue is 2SDs more severe than the mean of the general population. Our next step is to evaluate the clinical usefulness of the PROMIS FIB item bank and short forms in various clinical populations.

We used fatigue as an example of how the applications of a comprehensive item bank can be easily incorporated into the busy rehabilitation settings without increasing burden to patients and clinicians. Given its brief-yet-precise and easily administration characteristics, these applications can lead to early fatigue identification and further timely intervention. These applications are also applicable to other health-related quality of life and symptoms seen in rehabilitation settings and have drawn attention in the rehabilitation outcome measurement field and at the federal government level.

Study Limitation

Additional work is needed to evaluate the use of the FIB with a neurological sample, especially in the context of clinical trials. Subsequent to the PROMIS initiative, National Institute of Neurological Disorders and Stroke initiated an effort to develop item banks for people with neurological conditions, Neuro-QOL.[54] Work on that project is in progress. In addition, item banks for traumatic brain injury (TBI-QOL) and spinal cord injury (SCI-QOL) have been developed to enhance measurement in the rehabilitation field.[48–49]

Conclusion

In conclusion, PROMIS FIB is a psychometrically sound measurement tool. CAT and short-forms derived from the PROMIS FIB can reliably estimate fatigue reported by the US general population. Applications of the PROMIS FIB are available for use in rehabilitation settings.

Acknowledgments

Funding Source: The Patient-Reported Outcomes Measurement Information System (PROMIS) 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 have included Deborah Ader, PhD, Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Laura Lee Johnson, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Susana Serrate-Sztein, MD, and James Witter, MD, 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.

I certify that no party having a direct interest in the results of the research supporting this article has or will confer a benefit on me or on any organization with which I am associated AND, if applicable, I certify that all financial and material support for this research (eg, NIH or NHS grants) and work are clearly identified in the title page of the manuscript.

Abbreviations

FIB	Fatigue item bank
PROMIS	Patient Reported Outcomes Measurement Information System
CAT	Computerized Adaptive Testing
IRT	Item Response Theory
PRO	Patient-reported outcomes
FACIT - F	Functional Assessment of Chronic Illness Therapy - Fatigue
NIH	National Institute of Health

CFA	Confirmatory factor analysis
RMSEA	Root mean squared error of approximation
CFI	Comparative fit index
DIF	Differential item functioning
MH	Mantel-Haenszel
IF	Information function
EAP	Expected A Posteriori
MPWI	Maximum posterior weighted information

References

1. Davis KM, Cella D. Assessing quality of life in oncology clinical practice: A review of barriers and critical success factors. Journal of Clinical Outcomes Management. 2002; 9(6):327–332.
2. Bode RK, Lai JS, Cella D, Heinemann AW. Issues in the development of an item bank. Archives of Physical Medicine and Rehabilitation. 2003; 84 Suppl 2(4):S52–S60. [PubMed: 12692772]
3. Choppin B. An item bank using sample-free calibration. Nature. 1968; 219(156):870–872. [PubMed: 5673356]
4. McArthur DL, Choppin B. Computerized diagnostic testing. Journal of Educational Measurement. 1984; 21:391–397.
5. Wright BD, Bell SR. Item banks: What, why, how. Journal of Educational Measurement. 1984; 21(4):331–345.
6. Lai JS, Cella D, Dineen K, Von Roenn J, Gershon R. An item bank was created to improve the measurement of cancer-related fatigue. Journal of Clinical Epidemiology. 2005; 58(2):190–197. [PubMed: 15680754]
7. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Quality of Life Research. 2003; 12(5):485–501. [PubMed: 13677494]
8. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Quality of Life Research. 2007; 16(Suppl 1): 133–141. [PubMed: 17401637]
9. Gershon R, Cella D, Dineen K, Rosenbloom S, Peterman A, Lai JS. Item response theory and health-related quality of life in cancer. Expert Review of Pharmacoeconomics & Outcomes Research. 2003; 3(6):783–791. [PubMed: 19807355]
10. Weiss DJ, Kingsbury G. Application of computerized adaptive testing to educational problems. Journal of Educational Measurement. 1984; 21(4):361–375.
11. Hilberink SR, Roebroeck ME, Nieuwstraten W, Jalink L, Verheijden JM, Stam HJ. Health issues in young adults with cerebral palsy: towards a life-span perspective. Journal of Rehabilitation Medicine. 2007; 39(8):605–611. [PubMed: 17896051]
12. Bartels MN. Fatigue in cardiopulmonary disease. Physical Medicine and Rehabilitation Clinics of North America. 2009; 20(2):389–404. [PubMed: 19389619]
13. Pan JC, Bressler DN. Fatigue in rheumatologic diseases. Physical Medicine and Rehabilitation Clinics of North America. 2009; 20(2):373–387. [PubMed: 19389618]
14. Levine J, Greenwald BD. Fatigue in Parkinson disease, stroke, and traumatic brain injury. Physical Medicine and Rehabilitation Clinics of North America. 2009; 20(2):347–361. [PubMed: 19389616]
15. Shah A. Fatigue in multiple sclerosis. Physical Medicine and Rehabilitation Clinics of North America. 2009; 20(2):363–372. [PubMed: 19389617]

16. World Health Organization. Cancer pain relief and palliative care. Report of a WHO Expert Committee. World Health Organization Technical Report Series. 1990; 804:1–75. [PubMed: 1702248]

17. Cella D, Davis K, Breitbart W, Curt G. Cancer-related fatigue: Prevalence of proposed diagnostic criteria in a United States sample of cancer survivors. Journal of Clinical Oncology. 2001; 19(14): 3385–3391. [PubMed: 11454886]

18. Hjermstad MJ, Fayers PM, Bjordal K, Kaasa S. Health-related quality of life in the general Norwegian population assessed by the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire: the QLQ=C30 (+ 3). Journal of Clinical Oncology. 1998; 16(3):1188–1196. [PubMed: 9508207]

19. Stone P, Richardson A, Ream E, Smith AG, Kerr DJ, Kearney N. Cancer-related fatigue: Inevitable, unimportant and untreatable? Results of a multi-centre patient survey. Cancer Fatigue Forum. Annals of Oncology. 2000; 11(8):971–975. [PubMed: 11038033]

20. North American Nursing Diagnosis Association. Nursing diagnoses: Definition and Classification, 1997–1998. Philadelphia, PA: McGraw-Hill; 1996.

21. Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E. Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. Journal of Pain and Symptom Management. 1997; 13(2):63–74. [PubMed: 9095563]

22. Cella D, Yount S, Sorensen M, Chartash E, Sengupta N, Grober J. Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. Journal of Rheumatology. 2005; 32:811–819. [PubMed: 15868614]

23. Daut RL, Cleeland CS, Flanery RC. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain. 1983; 17(2):197–210. [PubMed: 6646795]

24. Piper BF, Dibble SL, Dodd MJ, Weiss MC, Slaughter RE, Paul SM. The revised Piper Fatigue Scale: Psychometric Evaluation in Women with Breast Cancer. Oncology Nursing Forum. 1998; 25(4):677–684. [PubMed: 9599351]

25. Smets EMA, Garssen B, Bonke B, DeHaes JCJM. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. Journal of Psychosomatic Research. 1995; 39:315–325. [PubMed: 7636775]

26. Stein KD, Martin SC, Hann DM, Jacobsen PB. A multidimensional measure of fatigue for use with cancer patients. Cancer Practice. 1998; 6(3):143–152. [PubMed: 9652245]

27. Lai JS, Crane PK, Cella D. Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. Quality of Life Research. 2006; 15(7):1179–1190. [PubMed: 17001438]

28. DeWalt DA, Rothrock N, Yount S, Stone AA. PROMIS Cooperative Group. Evaluation of Item Candidates: The PROMIS Qualitative Item Review. Medical Care. 2007; 45 Suppl 1(5):S12–S21. [PubMed: 17443114]

29. Ware JE Jr, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual Framework and Item Selection. Medical Care. 1992; 30(6):473–483. [PubMed: 1593914]

30. Molenaar IW, Sikkema K. MSP for Windows (Version 5). 2000

31. Hardouin JB. Manual for the SAS macro-programs LoevH and MSP and the Stata modules LoevH and MSP. Available at: http://www.freeirt.org/index.php?file=database/detailpgm.php&cond=idprogram=12 Accessed.

32. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6(1):1–55.

33. McDonald, RP. Test Theory: A unified treatment. Mahwah, NJ: Lawrence Earlbaum Associates, Inc.; 1999.

34. 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. Quality of Life Research. 2009; 18(4):447–460. [PubMed: 19294529]

35. Gibbons R, Hedeker D. Full-information item bi-factor analysis. Psychometrika. 1992; 57(3):423–436.

36. Lai JS, Butt Z, Wagner L, et al. Evaluating the dimensionality of perceived cognitive function. Journal of Pain and Symptom Management. 2009; 37(6):982–995. [PubMed: 19500722]

37. Lai JS, Cella D, Crane P. Cancer related fatigue is sufficiently unidimensional for applications requiring unidimensionality. Quality of Life Research. 2005; 14(9):1990–1990.

38. Orlando M, Thissen D. Further examination of the performance of S-X^2, an item fit index for dichotomous item response theory models. Applied Psychological Measurement. 2003; 27:289–298.

39. Zumbo, BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.

40. Mantel N. Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association. 1963; 58:690–700.

41. Zwick R, Thayer DT. Evaluating the magnitude of differential item functioning in polytomous items. Journal of Educational and Behavioral Statistics. 1996; 21(3):187–201.

42. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute. 1959; 22(4):719–748. [PubMed: 13655060]

43. Samejima, F.; van der Liden, WJ.; Hambleton, R. Handbook of modern item response theory. New York, New York: Springer; 1996. The graded response model; p. 85-100.

44. 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). Medical Care. 2007; 45 Suppl 1(5):S22–S31. [PubMed: 17443115]

45. Bock, RD.; Toit, Md. A brief history of Item Response Theory. In: Du Toit, M., editor. IRT from SSI: BIOLOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood, IL: Scientific Software International, Inc.; 2003.

46. Timminga, E.; Adema, JJ. Test construction from item banks. In: Fischer, GH.; Molenaar, IW., editors. Rasch models: Foundations, recent developments, and applications. New York: Springer-Verlag, New York, Inc; 1995. p. 111-130.

47. Baker, FB. The basics of Item Response Theory. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation; 2001.

48. Garcia SF, Cella D, Clauser SB, et al. Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. Journal of Clinical Oncology. 2007; 25(32):5106–5112. [PubMed: 17991929]

49. Choi S. Firestar: Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models. Applied Psychological Measurement. 2009; 33(8):644–645.

50. van der Linden, WJ.; Pashley, P. Item Selection and Ability Estimation in Adaptive Testing. In: van der Linden, WJ.; Glas, CAW., editors. Computerized Adaptive Testing: Theory and Practice. Boston: Kluwer Academic; 2000. p. 1-25.

51. Choi SW, Swartz RJ. Comparison of CAT item selection criteria for polytomous items. Applied Psychological Measurement. 2009; 33(6):419–440. [PubMed: 20011456]

52. Velikova G, Wright P, Smith AB, et al. Self-reported quality of life of individual cancer patients: Concordance of results with disease course and medical records. Journal of Clinical Oncology. 2001; 19(7):2064–2073. [PubMed: 11283140]

53. Smith E, Lai JS, Cella D. Building a measure of fatigue: the functional assessment of chronic illness therapy fatigue scale. Archives of Physical Medicine and Rehabilitation. 2010; 2(5):359–363.

54. Cella D, Victorson D, Nowinski C, Peterman A, Miller DM. The Neuro-QOL project: Using Multiple Methods to Develop a HRQOL Measurement Platform to be Used in Clinical Research Across Neurological Conditions. Quality of Life Research. 2006; A-14:1353.

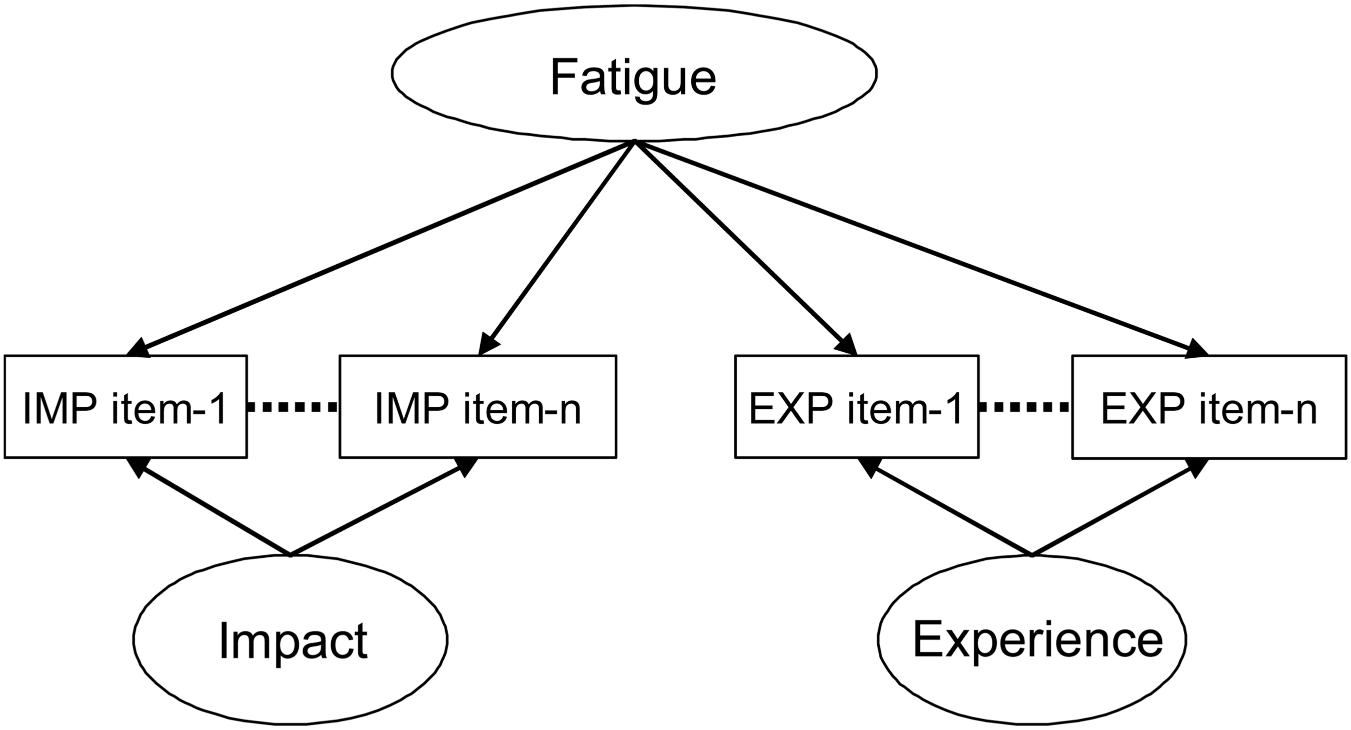

Figure 1.
Hypothesized Model Used for Bi-factor Analysis
NOTE:

- General factor is defined as “overall fatigue” and two local factors are “fatigue impact” and “fatigue experience”.
- IMP item-n: items measure “fatigue impact”
- EXP item-n: items measure “fatigue experience”
- CFI=0.911; TLI=0.996; RMSEA=0.100

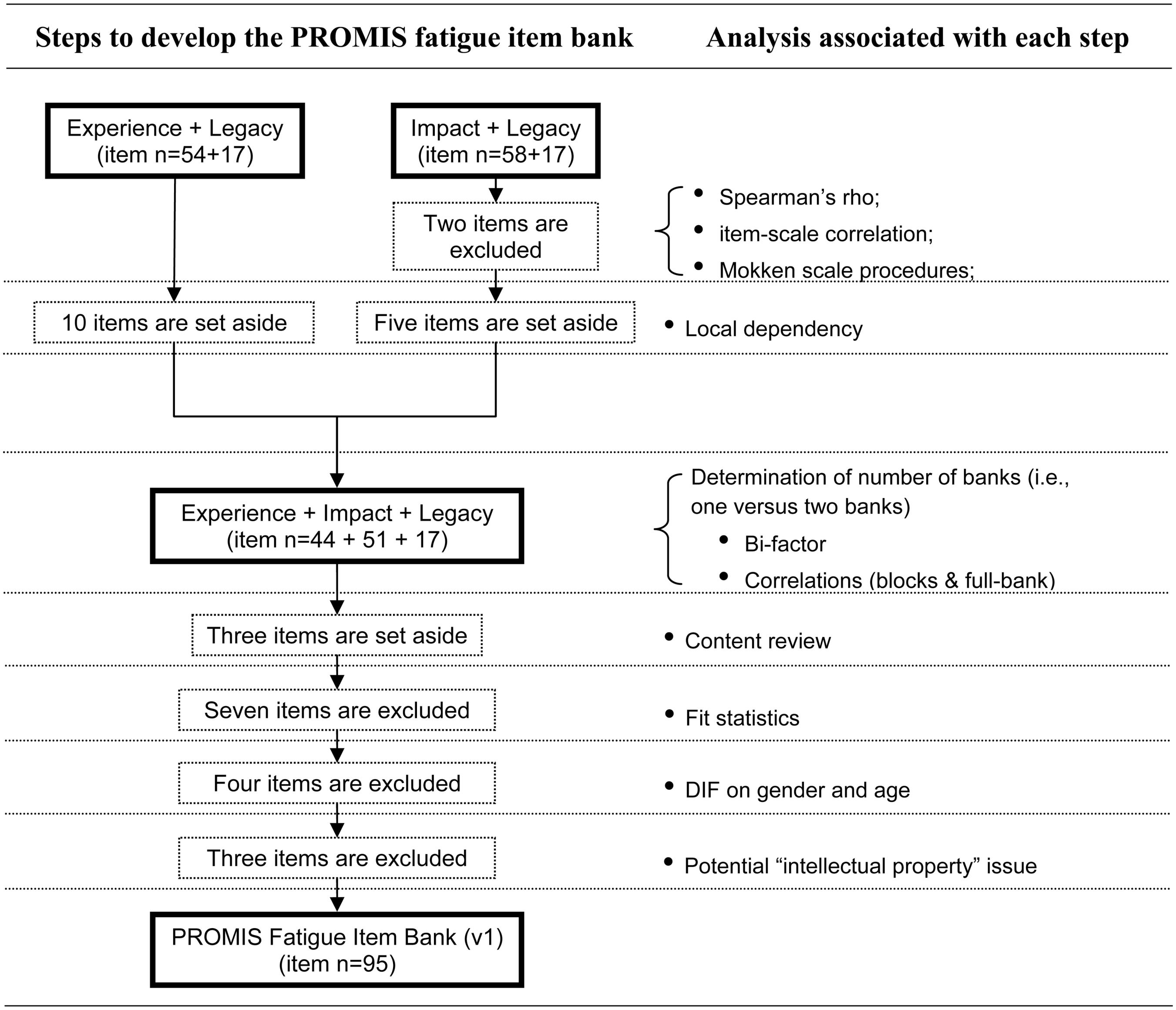
Steps to develop the PROMIS fatigue item bank
Analysis associated with each step
Experience + Legacy (item n=54+17)
Impact + Legacy (item n=58+17)
Two items are excluded
Spearman's rho;
item-scale correlation;
Mokken scale procedures;
10 items are set aside
Five items are set aside
Local dependency
Experience + Impact + Legacy (item n=44 + 51 + 17)
Determination of number of banks (i.e., one versus two banks)
Bi-factor
Correlations (blocks & full-bank)
Three items are set aside
Content review
Seven items are excluded
Fit statistics
Four items are excluded
DIF on gender and age
Three items are excluded
Potential "intellectual property" issue
PROMIS Fatigue Item Bank (v1) (item n=95)

Figure 2.

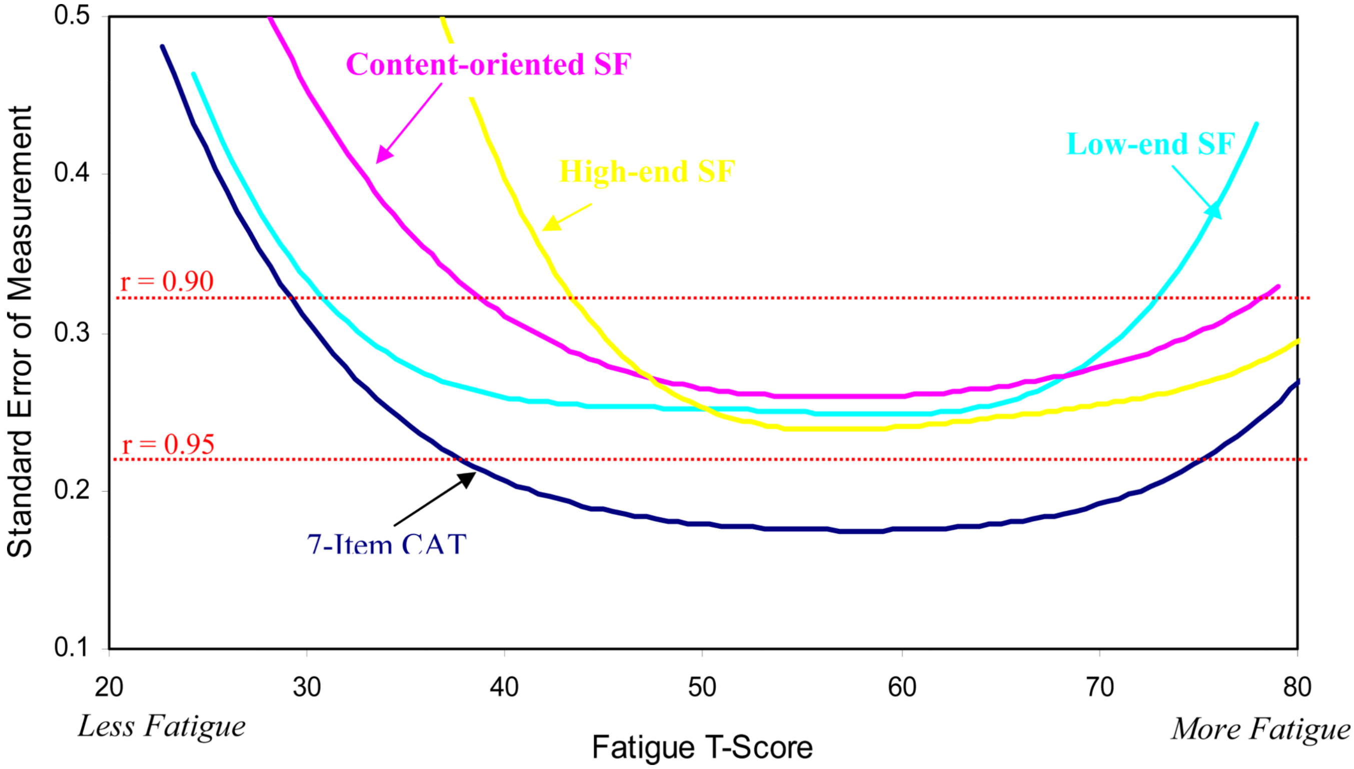

Figure 3.
Comparisons of Short-forms and CAT

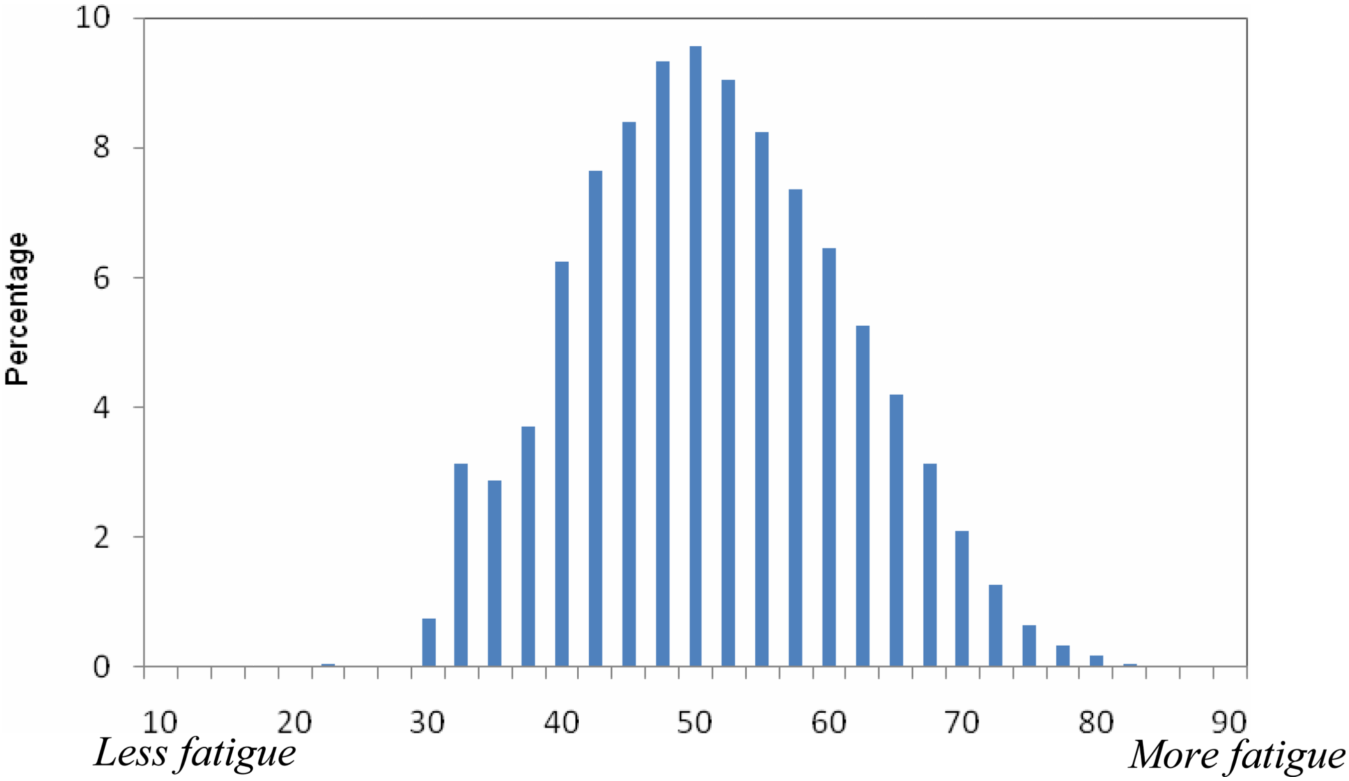

Figure 4.
Distribution of the Sample Fatigue Scores (in T-score).

Table 1

Item Stems of the Example PROMIS Fatigue Short-Forms

	Content-Oriented	High-End (Severe Fatigue)	Low-End (Mild Fatigue)
1	How often did you feel tired?	How often were you too tired to watch television?	How often were you sluggish?
2	How often did you experience extreme exhaustion?	How often did your fatigue make it difficult to make decisions?	How fatigued were you when your fatigue was at its worst?
3	How often did you run out of energy?	How often was it an effort to carry on a conversation because of your fatigue?	How often were you energetic?
4	How often did your fatigue limit you at work (include work at home)?	How often were you too tired to socialize with your family?	How tired did you feel on average?
5	How often were you too tired to think clearly?	How hard was it for you to carry on a conversation because of your fatigue?	How fatigued were you on the day you felt most fatigued?
6	How often were you too tired to take a bath or shower?	To what degree did your fatigue make it difficult to make decisions?	How energetic were you on average?
7	How often did you have enough energy to exercise strenuously?	How fatigued were you on the day you felt least fatigued?	How often did you find yourself getting tired easily?