Published in final edited form as:
Qual Life Res. 2013 November ; 22(9): . doi:10.1007/s11136-013-0357-1.

Development and Psychometric Properties of the PROMIS® Pediatric Fatigue Item Banks

Jin-Shei Lai, PhD, OTR/L[1], **Brian D. Stucky, PhD**[2], **David Thissen, PhD**[3], **James W. Varni, PhD**[4], **Esi Morgan DeWitt, MD**[5], **Debra E. Irwin, PhD**[6], **Karin B. Yeatts, PhD**[6], and **Darren A. DeWalt, MD, MPH**[7]

[1]Department of Medical Social Sciences and Pediatrics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[2]RAND Corporation, Santa Monica, CA, USA

[3]Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[4]Department of Pediatrics, College of Medicine, Department of Landscape Architecture and Urban Planning, College of Architecture, Texas A&M University, College Station, TX, USA

[5]Department of Pediatrics, Division of Rheumatology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[6]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[7]Division of General Medicine and Clinical Epidemiology, Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Abstract

PURPOSE—This paper reports on the development and psychometric properties of self-reported pediatric fatigue item banks as part of the Patient Reported Outcomes Measurement Information System (PROMIS).

METHODS—Candidate items were developed by using PROMIS qualitative methodology. The resulting 39 items (25 tiredness- and 14 energy-related) were field tested in a sample that included 3,048 participants aged 8–17 years. We used confirmatory factor analysis (CFA) to evaluate dimensionality, differential item functioning (DIF) analysis to evaluate parameter stability between genders and by age; we examined residual correlations to evaluate local dependence (LD) among items, and estimated the parameters of item response theory (IRT) models.

RESULTS—Of 3,048 participants, 48% were males, 60% were white and 23% had at least one chronic condition. CFA results suggest two moderately correlated factors. Two items were removed due to high LD, and three due to gender-based DIF. Two item banks were calibrated separately using IRT: Tired and (Lack of) Energy, which consisted of 23 and 11 items, respectively; ten- and 8-item short-forms were created.

CONCLUSION—The PROMIS assessment of self-reported fatigue in pediatrics includes two item banks: Tired and (Lack of) Energy. Both demonstrated satisfactory psychometric properties and can be used for research settings.

Corresponding Author: Jin-Shei Lai, Ph.D., OTR/L, Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 633 St Clair Street, #19-039, Chicago, IL 60611, TEL: 312-503-3370, FAX: 312-503-9800, js-lai@northwestern.edu.

Keywords

PROMIS; Fatigue; Children; Item Response Theory; Health-related Quality of Life; Patient Reported Outcomes

INTRODUCTION

Fatigue is defined as an overwhelming and sustained sense of exhaustion that decreases one's capacity for physical and mental work.[1] It is one of the most universal experiences for children with a number of pediatric chronic health conditions such as juvenile idiopathic arthritis,[2] diabetes,[3] inflammatory bowel disease,[4]multiple sclerosis,[5] fibromyalgia,[6] epilepsy,[7] and cancer.[8; 9] Across these conditions, fatigue is significantly related to difficulties in sleep, cognition, physical functioning, emotional functioning, appetite, academic achievement, and overall quality of life.[7–12] Yet fatigue is also a common experience among children without chronic conditions with prevalence ranging 25–40%.[13–15]

The etiology of fatigue is poorly understood, may vary across disease conditions, and is likely multi-factorial. It can be caused directly by disease-related features such as anemia, inflammation, and stress, and secondarily by treatment effects. Consequently, fatigue assessment is controversial and laboratory tests provide explanations of some but not all fatigue, such as hemoglobin values for anemic fatigue. While clinicians typically use patient- or observer-reported questionnaires to evaluate fatigue status, there is considerable variability among methods of assessment. In our review of the OVID MEDLINE database from 1950 to November 2012, 268 manuscripts were initially identified using the keywords "fatigue', "cancer or brain tumor", and "ped* or child*." Of these 268 manuscripts, only 78 quantified self-reported fatigue. The remainder comprised literature reviews (n=55), qualitative evaluations of fatigue (n=21), or "others" (e.g., not pediatric samples, not written in English or non-cancer samples). Among the 78 manuscripts that used quantitative approaches, many (42 of 78) measured fatigue in an unstructured manner such as using general statements from patients/parents, adverse event documentation, chart review, selected questionnaire items from the literature, or single-item measures; some failed to mention how fatigue was assessed. These findings were similar when unspecified health conditions (as opposed to cancer) were used. Eddy and Cruz[16] identified 11 articles in their search for fatigue and quality of life for children with chronic health problems in general. Of these 11 articles, seven were cancer-related, two involved juvenile arthritis, and one considered both rheumatologic diagnosis and epilepsy. Most of these studies used either qualitative (n=3) or descriptive (n=5) designs. Both systematic reviews indicated that well-accepted measurement tools are needed. This lack of any standard measurement tool contributes to wide ranges of estimated prevalence not only across various diseases/conditions but also within specific diseases/conditions.

Several pediatric instruments have been developed to measure fatigue such as the PedsQL™ Multidimensional Fatigue Inventory,[17] Fatigue Scale,[18] Memorial Symptom Assessment Scale (MSAS) Pediatric 10–18,[19] Chalder Fatigue Scale,[20; 21] and pediatric Functional Assessment of Chronic Illness-Fatigue (pedsFACIT-F).[22] Most of these scales were developed using classical test theory, although the pedsFACIT-F was developed using Rasch [23] analysis. The latter was initially designated for children with cancer; its generalizability to children with or without other chronic conditions has not yet been evaluated.

Given the high prevalence of fatigue in children, a psychometrically sound measurement tool is important to facilitate a better understanding and quantification of the impact of fatigue on daily functioning of pediatric populations across a variety of chronic health conditions. The Patient Reported Outcomes Measurement and Information System, PROMIS®, originated as a National Institutes of Health (NIH) Common Fund that has developed item response theory (IRT)-based item banks, collections of items which measure a single domain of health, to measure patient-reported symptoms and other aspects of health-related quality of life across various conditions and disease populations.[24; 25] Unlike traditional classical test theory that weights items equally, IRT calibrates items and patients onto the same latent trait or continuum (in this study, fatigue) according to the degree of fatigue a patient experiences and the severity of fatigue an item measures. The advantage of an item bank is that patients do not need to take the same items in order to have comparable scores. Thus, a well-calibrated item bank can facilitate brief-yet-precise assessment by enabling static or dynamic adaptive testing via fixed-length short-forms or computerized adaptive testing (CAT), respectively.[26]

The pediatric fatigue item bank was among the item banks developed during the PROMIS Wave I (2004–2009) efforts. This study reports on the development and psychometric properties of the PROMIS pediatric fatigue item banks and potential clinical research applications.

METHODS

Participants and Procedures

Institutional Review Boards at all participating sites approved this study. The same sampling strategy was used for all PROMIS pediatric item banks developed during 2004–2009 (i.e., pain interference, fatigue, physical function-mobility, physical function-upper extremity, anxiety, depressive symptoms, peer relationships, and asthma symptoms) and are described by Irwin et al.[27] In brief, candidate pediatric fatigue items were administered to a racially diverse cohort of 3,048 children aged 8–17. Approximately 23% of this sample had a chronic medical condition and were recruited from hospital-based outpatient general pediatrics clinics, subspecialty clinics (Pulmonology, Allergy, Gastroenterology, Rheumatology, Nephrology, Obesity, and Endocrinology) and in public school settings (the Chapel Hill-Carrboro Public School System, North Carolina) between January 2007 and May 2008 in North Carolina and Texas (see Table 1 for participant characteristics).

North Carolina and Texas were chosen as recruitment sites because of the diversity of cultural experience and population characteristics that existed in those areas. Inclusion criteria were: ages 8 to 17 years old; able to speak and read English; and able to complete questions using a computer. After informed consent and assent forms were obtained, children completed PROMIS pediatric candidate items.

Generation of the Pediatric Fatigue Item Pool

Items included in the PROMIS pediatric fatigue item pool were generated using a methodology consistent within the PROMIS Network,[28] including identification of existing items, item classification and selection, item review and revision, focus group input on domain coverage, cognitive interviews for individual items, and final revision before field testing. Identification of items refers to the systematic search for existing items in currently available pediatric scales. After the first round of review by the research team, the initial item pool consisted of 68 fatigue items in the content areas of physical function, emotional functioning, social functioning, cognitive functioning and sleep. Expert item review and revision was conducted by trained professionals who reviewed the wording of each item and revised as appropriate for conventions adopted by the PROMIS network [25; 28] Simultaneously, focus groups were used to confirm domain definitions, and to identify new areas of item development for future PROMIS item banks.[29] Cognitive interviews were used to examine and refine wording of individual items.[30] As a result, 39 items were retained and field tested. Items were written in the past tense with a seven-day recall period and utilized a 5-point rating scale (never, almost never, sometimes, often, almost always).

These items were randomly assigned to 4 different test forms as shown on Table 2; each test form also included items from other PROMIS domains. Each child was randomly assigned to only one testing form, and each item was administered to at least 754 children. Item banks were subsequently developed by combining selected items across test form. This linking technique, "common population linking," is based on calibrating items from multiple test forms that have been randomly assigned to a common population.[31] This strategy was designed to minimize respondent burden and to allow for the following: 1) evaluation of domain factor structure, 2) testing stability of measurement properties between sub-sample - differential item function (DIF), 3) evaluation for local dependence (LD), and 4) IRT calibrations.

Statistical and psychometric methods

The study population was characterized by descriptive statistics. We followed a set of standard PROMIS procedures for psychometric item analyses. [32; 33] Data quality was verified and analyses were conducted to ensure that IRT model assumptions were met. Construct unidimensionality was assessed with confirmatory factor analysis (CFA) of the inter-item polychoric correlation matrices using the WLSMV algorithm in the computer program Mplus.[34] Items were considered unidimensional when the comparable fit index (CFI) > 0.90, Tucker-Lewis Index (TLI) > 0.90, and Root Mean Square Error Approximation (RMSEA) <=0.08. Local dependence was investigated using modification indices and by examining residual correlations between item pairs.[35] If LD was identified between multiple items on the same factor, only one of the items was selected from the subset to remain in the item bank and the others were set aside.

Items that met the unidimensionality criterion were subsequently calibrated using Samejima's Graded Response Model (GRM)[36; 37] as implemented in the computer program Multilog.[38] For each item the GRM estimates a slope or discrimination parameter (a), which indicates the degree of association between the item responses and the underlying construct and four thresholds (b_k) (for five category items) that reflect the degree of fatigue where the most probable response occurs in a given category or higher. The item-fit of the IRT model was evaluated using $S\text{-}X^2$ statistic[39] as implemented in the SAS macro IRTFIT for polytomous response items.[40]

DIF between genders and ages (ages 8–11 versus 12–17) was evaluated using IRT likelihood-ratio tests as implemented in IRTLRDIF.[41; 42] The Benjamini-Hochberg procedure was used to make inferential decisions in the context of the multiple comparisons.[43] The magnitude of the effect size was evaluated graphically using methods outlined by Steinberg and Thissen [44] for items with significant DIF. The research team considered item inclusion or exclusion by reviewing the analytic results, graphical illustrations, item content, and clinical relevance. *For improved communication to end-users, the standardized IRT-metric was then transformed to a T-score metric (mean=50, standard deviation=10), in which 50 represents the mean of the general population sample that this study tested.* IRT-based item information functions were computed for each item. Information reflects the degree of precision level along the fatigue measurement continuum. Higher information implies less measurement error and therefore better precision.[22] We developed suggested short-forms by selecting the most informative items at the sample mean (i.e., T-score =50) and considering item content representative of the bank.

RESULTS

Preliminary analyses using both confirmatory and exploratory factor analyses (EFA) strongly suggested the presence of two correlated dimensions (*Lack of) Energy* and *Tired. Results of these analyses are available upon request.* Specifically, (*Lack of) Energy* items were positively worded including the phrase of "I had enough energy …"; while *Tired* items were negatively worded such as "I was too tired to …" or "Being tired made it hard for me to …". Items were scored so that higher scores indicate being tired or lacking energy. Subsequent two-factor models indicated that inter-factor correlations between (Lack of) Energy and Tired ranged from 0.49 to 0.74 across four testing forms. Given these moderate correlations, (Lack of) Energy and Tired were calibrated separately.

Table 2 provides factor loadings and residual correlations for each form. All forms showed acceptable fit indices: CFI = 0.99, TLI = 0.99, RMSEA = 0.05; CFI = 0.99, TLI = 0.99, RMSEA = 0.04; CFI = 0.94, TLI = 0.97, RMSEA = 0.08; CFI = 0.98, TLI = 0.99, RMSEA = 0.04, for forms 1–4, respectively. In Form 1, the two items mentioning "energy" did not suggest a second factor; instead, "I had enough energy to eat" was locally dependent with "I was too tired to eat" (residual correlation=0.42). After setting aside "I had enough energy to eat", the only negatively-worded "energy" item, "I did not have much energy", was also set aside. The other instance of local dependence occurs on Form 4 between the items "I had enough energy to focus on my work" and "I was so tired it was hard for me to focus on my work" (residual correlation=0.39). However, because "tired" and "energy" items were to be calibrated separately, both items remained on their respective scales for calibration. As a result, 37 of the original 39 items were calibrated.

Tables 3a and 3b provide the IRT parameters, item fit statistics ($S\text{-}X^2$), and DIF statistics ($LR\ X^2$) for the separately calibrated *Tired* and *(Lack of) Energy* domains. Slope parameters of *Tired* items (i.e., discrimination power) ranged from 0.91 (*I was too tired to watch television*) to 1.90 (*I was too tired to enjoy the things I like to do*). For *(Lack of) Energy* items, it ranged from 1.09 (I had enough energy to read) to 2.58 (*I had enough energy to do the things I like to do*). Listed in footnotes to Tables 3a and 3b are three items set aside due to gender DIF: two items from the *Tired* domain, and a single item from the *(Lack of) Energy* domain. Two additional items from the *(Lack of) Energy* dimension showed statistically significant DIF, but the effect sizes between males and females were relatively small. For these items, graphical illustrations shown in Figure 1 indicate that the expected value of the item response on the 0–4 response scale is much less than half a point across the fatigue range for both items, and therefore both items were retained. In terms of age DIF, 16 of the 25 *Tired* items exhibited DIF and three additional items exhibited DIF at the uncorrected 0.05 level. Specifically, items discriminated among individuals in the older age group better than those in the younger group. This suggests the concept of "tired" might not be the same between age groups. However, no items were removed due to age DIF for two reasons: 1) items performed well within each age group; and 2) removing more than 50% of items due to DIF may change the meaning of the underlying construct. A similar conclusion was drawn for (*Lack of) Energy* items, among which 6 of the 11 items exhibited significant age DIF. The final calibrated *Tired* and (*Lack of) Energy* item banks contained 23 and 11 items, respectively.

Items with the highest magnitude of information around the mean were selected to construct short-forms; 10- and 8-item short-forms were created for *Tired* and (*Lack of) Energy*, respectively (see Tables 3a and 3b). Figures 2 and 3 illustrate item bank and short-form information functions for the (*Lack of) Energy* and *Tired* domains. Test information is the expected value of the inverse of the squared standard error of measurement, such that a standard error of measurement of approximately 0.45 (or 4.5 on a *T*-score metric) is associated with a test information value of 5, which corresponds to a reliability coefficient of 0.80. The 8-item (*Lack of) Energy* short-form has information values greater than 5 between scores of approximately 40 to 80 (which covers approximately 84% of the population). The 10-item *Tired* short-form has information values greater than 5 between scores of about 35 to 85 (which covers approximately 93% of the population). A summed score to IRT score translation table[41] is provided that allows the user to take advantage of IRT scoring without conducting an IRT analysis (see Table 4).

Because the items are informative in generally the same locations, a CAT may not dramatically improve the efficiency of the item bank beyond the performance of these short-forms. Nonetheless, PROMIS Assessment Center^{SM} (http://www.assessmentcenter.net/) contains the calibrated item banks and allows the user to select and administer items as a CAT.

DISCUSSION

The present study produced two calibrated item banks to measure the domain of fatigue from the pediatric perspective, *Tired* and (*Lack of) Energy*. Items from the *Tired* and *(Lack of) Energy* item banks reflect children's own perception of fatigue and its impact on their daily lives. These two item banks demonstrated acceptable psychometric properties, including unidimensionality, consistent measurement properties between males and females evaluated by DIF, and satisfactory fit statistics.

The concepts and advantages of using IRT in health-related outcomes research have been documented in detail elsewhere.[24; 26; 35; 45; 46] Briefly, the major advantage of using IRT principles for patient-reported outcomes is that it enables adaptive testing, either dynamic via computerized adaptive testing (CAT) or static via multiple short-forms. Although CAT typically produces estimates with consistent precision along the measurement continuum due to its individually tailored item selection procedures, it requires computer access that might not be feasible for many clinical settings. Thus, short-forms can be valuable alternatives. In this paper, we demonstrate two examples of short-forms that are intended to be the most informative around the mean and would serve well for most studies. However, with access to the item bank, researchers may find utility in some other set of items more tailored to a study's needs. For example, a short-form targeting children with severe fatigue can be produced for clinical trials at baseline, and the one targeting moderate or mild fatigue, in anticipation of improvement, can be used for the exit assessment. In this scenario, scores from the different short-forms taken from a unidimensional item bank are comparable because the items are calibrated on the same measurement continuum. IRT-based scoring tables can be produced as shown on Table 4; thus, investigators can easily monitor changes in fatigue over time in a very brief-yet-precise manner.

Whether fatigue should be measured using a unidimensional or multidimensional approach is still debated. Recent studies[47; 48] indicated sufficient unidimensionality of fatigue in adult patients. We were unable to obtain evidence to support unidimensionality of fatigue in this study and decided to produce two separate calibrated item banks. Because the *Tired* bank appears most consistent with the adult fatigue item bank, we have selected the *Tired* bank as the pediatric fatigue item bank recommended for use. There may be applications for the *(Lack of) Energy* bank, but we anticipate that it will be used less frequently.

Parameters reported in this paper were based on a sample without pre-defined medical conditions. We took this approach not only because fatigue is a prevalent complaint among children but also because it can serve as a reference to enable fatigue comparisons across children with various chronic conditions. The latter is particularly important from a clinical perspective. As mentioned earlier, although the etiology of fatigue is not well-understood, the clinical manifestations are similar regardless of cause. We hope the PROMIS pediatric fatigue item banks can serve as a frame of reference, which can help investigators to determine the significance of fatigue and whether an intervention is necessary. To facilitate user-friendly interpretations, we converted the IRT-scaled scores into a T-score metric with mean =50 and standard deviation = 10. For example, a child reporting a fatigue score of 60 indicates that this child is one standard deviation more severe than our calibration population. This child is likely to endorse “sometimes” for item “I was too tired to do things outside” while children with a fatigue score=50 are likely to endorse “almost never” on this same item. In this scenario the respondent is not only more severe than 84% of general population but also exceeds a distribution-based clinically minimal important difference, which was recommended to be as small as 0.25 to 0.33 SD;[49; 50] Therefore, medical evaluation might be warranted. Yet determinations of cut-off scores based on clinical variables should be estimated in future studies for patients with different conditions. Future studies to validate the PROMIS pediatric fatigue item banks in clinical populations are necessary (and currently underway) to make this comparison valid and clinically meaningful.

In this study, age DIF was identified on more than 50% of items. To our knowledge this is the first large-scale study of pediatric fatigue to use IRT-based DIF analysis with respect to age. Our findings suggest that fatigue may change conceptually across childhood. That we found DIF does not mean that the scale is not appropriate for younger or older children, but it does mean that caution is advised comparing scores by age. Other studies evaluating fatigue in children with chronic conditions (e.g., cancer[22]) found stable parameter estimates across age groups, albeit with much smaller samples so the difference in the results may be due to differences in power. The study team decided not to remove items because of age DIF because that process would not have resulted in a useful bank of items; instead, we recommend using these items over shorter time frames and not in long-term longitudinal studies. Future studies are needed in order to further evaluate how fatigue conceptually changes according to age.

The PROMIS pediatric data collection is a large-scale study with sufficient statistical power, yet some limitations are noted. Given budgetary constraints, participants were not geographically stratified and did not match the composition of the US pediatric population. Instead, participants were drawn from two culturally diverse states, North Carolina and Texas. Though we believe our large sample size recruited from two diverse communities provides a useful population for norming, future studies using samples from other US pediatric populations are needed to evaluate item parameter stability. We randomly assigned items onto four separate forms and then randomly assigned participants to one of the four forms. This approach, known as common population linking, was taken to minimize participant burden. As shown in Table 1, the randomization was largely successful; differences in participant characteristics are consistent with random sampling error. However, this process comes with the limitation that we are unable to ensure that there are no local dependences across form. Studies involving the entire banks are in progress; those data will provide a check on the decisions that have been made here.

In conclusion, the PROMIS pediatric fatigue item banks are psychometrically sound measurement tools and are ready to be used in research settings. Clinical validation studies (currently ongoing) are warranted before use in clinical trials.

Acknowledgments

Funding:

This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1U01AR052181-01. Information on the Patient-Reported Outcomes Measurement Information System (PROMIS) can be found at http://nihroadmap.nih.gov.

References

1. North American Nursing Diagnosis Association. Nursing diagnoses: Definition and Classification, 1997–1998. Philadelphia, PA: McGraw-Hill; 1996.
2. Butbul-Aviel Y, Stremler R, Benseler SM, Cameron B, Laxer RM, Ota S, Schneider R, Spiegel L, Stinson JN, Tse SML, Feldman BM. Sleep and fatigue and the relationship to pain, disease activity and quality of life in juvenile idiopathic arthritis and juvenile dermatomyositis. Rheumatology. 2011; 50(11):2051–2060. [PubMed: 21873265]
3. Levy-Marchal C, Papoz L, de Beaufort C, Doutreix J, Froment V, Voirin J, Czernichow P. Clinical and laboratory features of type 1 diabetic children at the time of diagnosis. Diabetic Medicine. 1992; 9(3):279–284. [PubMed: 1576813]
4. Marcus SB, Strople JA, Neighbors K, Weissberg–Benchell J, Nelson SP, Limbers C, Varni JW, Alonso EM. Fatigue and Health-Related Quality of Life in Pediatric Inflammatory Bowel Disease. Clinical Gastroenterology and Hepatology. 2009; 7(5):554–561. [PubMed: 19418604]
5. Amato MP, Goretti B, Ghezzi A, Lori S, Zipoli V, Moiola L, Falautano M, De Caro MF, Viterbo R, Patti F, Vecchio R, Pozzilli C, Bianchi V, Roscio M, Martinelli V, Comi G, Portaccio E, Trojano M. Society FtMSSGotIN. Cognitive and psychosocial features in childhood and juvenile MS. Neurology. 2010; 75(13):1134–1140. [PubMed: 20876467]
6. Buskila D. Pediatric fibromyalgia. Rheumatic Diseases Clinics of North America. 2009; 35(2):253–261. [PubMed: 19647140]
7. Elliott IM, Lach L, Smith ML. I just want to be normal: A qualitative study exploring how children and adolescents view the impact of intractable epilepsy on their quality of life. Epilepsy and Behavior. 2005; 7(4):664–678. [PubMed: 16140594]
8. Wolfe J, Grier HE, Klar N, Levin SB, Ellenbogen JM, Salem-Schatz S, Emanuel EJ, Weeks JC. Symptoms and suffering at the end of life in children with cancer. The New England Journal of Medicine. 2000; 342(5):326–333. [PubMed: 10655532]
9. Jalmsell L, Kreicbergs U, Onelöv E, Steineck G, Henter JI. Symptoms Affecting Children With Malignancies During the Last Month of Life: A Nationwide Follow-up. Pediatrics. 2006; 117(4):1314. [PubMed: 16585329]
10. MacAllister WS, Christodoulou C, Troxell R, Milazzo M, Block P, Preston TE, Bender HA, Belman A, Krupp LB. Fatigue and quality of life in pediatric multiple sclerosis. Multiple Sclerosis. 2009; 15(12):1502–1508. [PubMed: 19965517]
11. Schanberg LE, Gil KM, Anthony KK, Yow E, Rochon J. Pain, stiffness, and fatigue in juvenile polyarticular arthritis: contemporaneous stressful events and mood as predictors. Arthritis and Rheumatism. 2005; 52(4):1196–1204. [PubMed: 15818661]
12. Meeske K, Katz ER, Palmer SN, Burwinkle T, Varni JW. Parent proxy-reported health-related quality of life and fatigue in pediatric patients diagnosed with brain tumors and acute lymphoblastic leukemia. Cancer. 2004; 101(9):2116–2125. [PubMed: 15389475]
13. Currie, C.; Hurrelmann, K.; Setterbulte, W.; Smith, R.; Todd, J. World Health Organization. Health and health behaviour among young people: health behaviour in school-aged children. Copenhagen, Denmark: World Health Organization Regional Office for Europe; 2000.
14. Ghandour RM, Overpeck MD, Huang ZJ, Kogan MD, Scheidt PC. Headache, Stomachache, Backache, and Morning Fatigue Among Adolescent Girls in the United States: Associations With Behavioral, Sociodemographic, and Environmental Factors. Archives of Pediatrics & Adolescent Medicine. 2004; 158(8):797. [PubMed: 15289254]

15. Viner RM, Clark C, Taylor SJC, Bhui K, Klineberg E, Head J, Booy R, Stansfeld SA. Longitudinal Risk Factors for Persistent Fatigue in Adolescents. Archives of Pediatrics and Adolescent Medicine. 2008; 162(5):469–475. [PubMed: 18458194]

16. Eddy L, Cruz M. The relationship between fatigue and quality of life in children with chronic health problems: A systematic review. Journal for Specialists in Pediatric Nursing. 2007; 12(2): 105–114. [PubMed: 17371554]

17. Varni JW, Burwinkle TM, Katz ER, Meeske K, Dickinson P. The PedsQL in pediatric cancer: Reliability and validity of the Pediatric Quality of Life Inventory Generic Core Scales, Multidimensional Fatigue Scale, and Cancer Module. Cancer. 2002; 94(7):2090–2106. [PubMed: 11932914]

18. Hinds PS, Hockenberry M, Tong X, Rai SN, Gattuso JS, McCarthy K, Pui CH, Srivastava DK. Validity and reliability of a new instrument to measure cancer-related fatigue in adolescents. Journal of Pain and Symptom Management. 2007; 34(6):607–618. [PubMed: 17629669]

19. Collins JJ, Byrnes ME, Dunkel IJ, Lapin J, Nadel T, Thaler HT, Polyak T, Rapkin B, Portenoy RK. The measurement of symptoms in children with cancer. Journal of Pain and Symptom Management. 2000; 19(5):363–377. [PubMed: 10869877]

20. Chalder T, Berelowitz G, Pawlikowska T, Watts L, Wessely S, Wright D, Wallace EP. Development of a fatigue scale. Journal of Psychosomatic Research. 1993; 37(2):147–153. [PubMed: 8463991]

21. Goligher EC, Pouchot J, Brant R, Kherani RB, Avina-Zubieta JA, Lacaille D, Lehman AJ, Ensworth S, Kopec J, Esdaile JM, Liang MH. Minimal clinically important difference for 7 measures of fatigue in patients with systemic lupus erythematosus. J Rheumatol. 2008; 35(4):635–642. [PubMed: 18322987]

22. Lai JS, Cella D, Kupst MJ, Holm S, Kelly ME, Bode RK, Goldman S. Measuring fatigue for children with cancer: development and validation of the pediatric Functional Assessment of Chronic Illness Therapy-Fatigue (pedsFACIT-F). Journal of Pediatric Hematology/Oncology. 2007; 29(7):471–479. [PubMed: 17609625]

23. Wright, BD.; Masters, GN. Rating scale analysis: Rasch measurement. Chicago: MESA Press; 1985.

24. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, Devellis R, Dewalt D, Fries JF, Gershon R, Hahn EA, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R, Lai JS. PROMIS Cooperative Group. The Patient Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology. 2010; 63(11): 1179–1194. [PubMed: 20685078]

25. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group During its First Two Years. Medical Care. 2007; 45(5 Suppl 1):S3–S11. [PubMed: 17443116]

26. 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]

27. Irwin DE, Stucky BD, Thissen D, Dewitt EM, Lai JS, Yeatts K, Varni JW, Dewalt DA. Sampling plan and patient characteristics of the PROMIS pediatrics large-scale survey. Quality of Life Research. 2010; 19(4):585–594. [PubMed: 20204706]

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(5 Suppl 1):S12–S21. [PubMed: 17443114]

29. Walsh T, Irwin D, Meier A, Varni J, DeWalt D. The use of focus groups in the development of the PROMIS pediatrics item bank. Quality of Life Research. 2008; 17(5):725–735. [PubMed: 18427951]

30. Irwin DE, Varni JW, Yeatts K, DeWalt DA. Cognitive interviewing methodology in the development of a pediatric item bank: a patient reported outcomes measurement information system (PROMIS) study. Health and Quality of Life Outcomes. 2009; 7(3)

31. Kolen, MJ.; Brennan, RL. Test equating, scaling, and linking : methods and practices. New York: Springer; 2004.

32. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai JS, Cella D. 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(5 Suppl 1):S22–S31. [PubMed: 17443115]

33. Yeatts KB, Stucky B, Thissen D, Irwin D, Varni JW, DeWitt EM, Lai JS, DeWalt DA. Construction of the Pediatric Asthma Impact Scale (PAIS) for the Patient-Reported Outcomes Measurement Information System (PROMIS). Journal of Asthma. 2010; 47(3):295–302. [PubMed: 20394514]

34. Joreskog, K.; Sorbom, D. LISREL 8.5. Lincolnwood, IL: Scientific Software International, Inc; 2003.

35. Hill CD, Edwards MC, Thissen D, Langer MM, Wirth RJ, Burwinkle TM, Varni JW. Practical Issues in the Application of Item Response Theory: A Demonstration Using Items From the Pediatric Quality of Life Inventory (PedsQL) 4.0 Generic Core Scales. Medical Care. 2007; 45(5 Suppl 1):S39–S47. [PubMed: 17443118]

36. Samejima, F. Psychometrika Monograph Supplement. 1969. Estimation of latent ability using a response pattern of graded scores.

37. Samejima, F. The graded response model. In: van der Linden, WJ.; Hambleton, R., editors. Handbook of modern item response theory. New York: Springer-Verlag; 1997. p. 85-100.

38. Du Toit, M. IRT from SSI : BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood, Ill: Scientific Software International; 2003.

39. 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.

40. Bjorner, JB.; Smith, KJ.; Edelen, MO.; Stone, C.; Thissen, D. IRTFIT: A Macro for Item Fit and Local Dependence Tests under IRT Models. Lincoln, RI: QualityMetric Incorporated; 2007.

41. Thissen, D. IRTLRDIF -Software for the computation of the statistics involved in item response theory likelihood-ratio test for differential item functioning (Version 2.0b). 2003.

42. Thissen, D.; Steinberg, L.; Wainer, H.; Holland, PW.; Wainer, H. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993. Detection of differential item functioning using the parameters of item response models; p. 63-113.

43. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57(1):289–300.

44. Steinberg L, Thissen D. Using Effect Sizes for Research Reporting: Examples Using Item Response Theory to Analyze Differential Item Functioning. Psychological Methods. 2006; 11(4): 402–415. [PubMed: 17154754]

45. Lai JS, Cella D, Choi SW, Junghaenel DU, Christodolou C, Gershon R, Stone A. How Item Banks and Their Application Can Influence Measurement Practice in Rehabilitation Medicine: A PROMIS Fatigue Item Bank Example. Archives of Physical Medicine and Rehabilitation. 2011; 92(10 Supplement):S20–S27. [PubMed: 21958919]

46. Lai JS, Butt Z, Zelko F, Cella D, Krull K, Kieran M, Goldman S. Development of a Parent-Report Cognitive Function Item Bank Using Item Response Theory and Exploration of its Clinical Utility in Computerized Adaptive Testing. Journal of Pediatric Psychology. 2011; 36(7):766–779. [PubMed: 21378106]

47. 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]

48. Cella D, Lai JS, Stone A. Self-reported fatigue: one dimension or more? Lessons from the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F) questionnaire. Supportive Care in Cancer. 2010; 19(9):1441–1450. [PubMed: 20706850]

49. Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of

Cancer Therapy (FACT) anemia and fatigue scales. Journal of Pain and Symptom Management. 2002; 24(6):547–561. [PubMed: 12551804]

50. Revicki D, Hays R, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology. 2008; 61(2):102–109. [PubMed: 18177782]

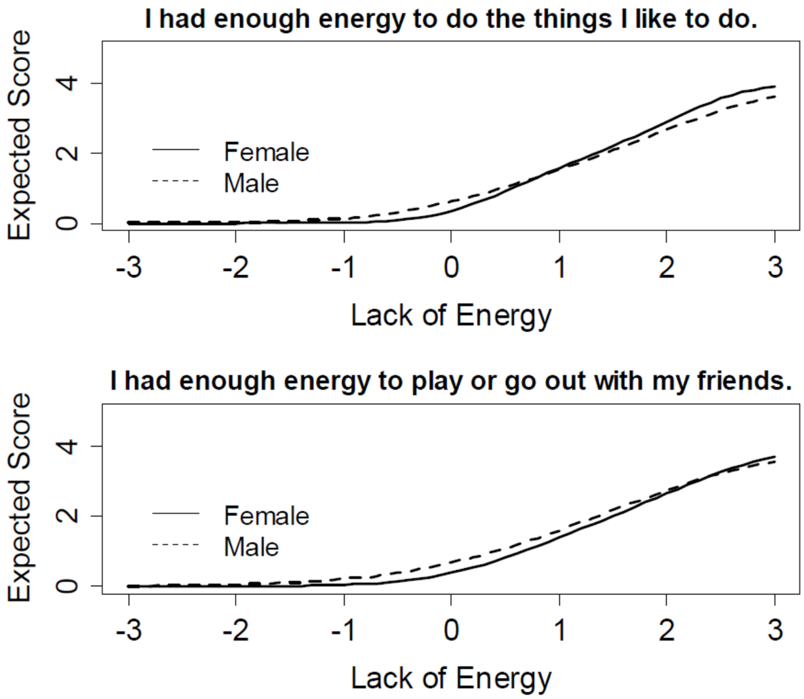

Figure 1.

Item characteristic curves between males and females for items demonstrating gender DIF. These two items were “I had enough energy to do the things I like to do” and “I had enough energy to play or go out with my friends”.

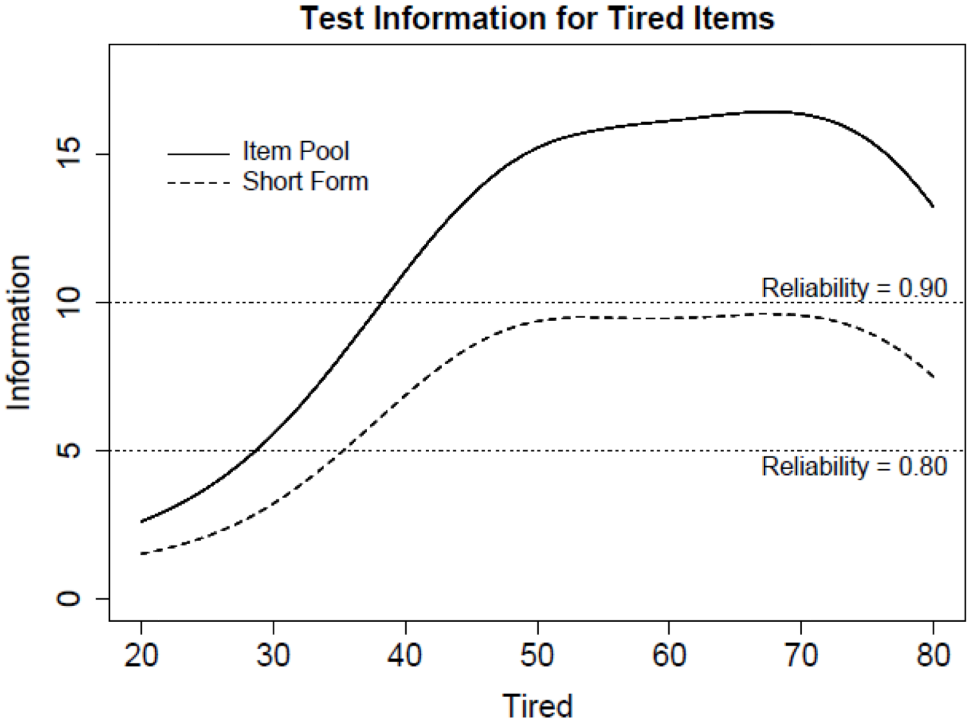

Figure 2.

Test information curves are displayed for the 11 *(Lack of) Energy* item pool, the 8 item short form, and the most informative 8 items at various score locations.

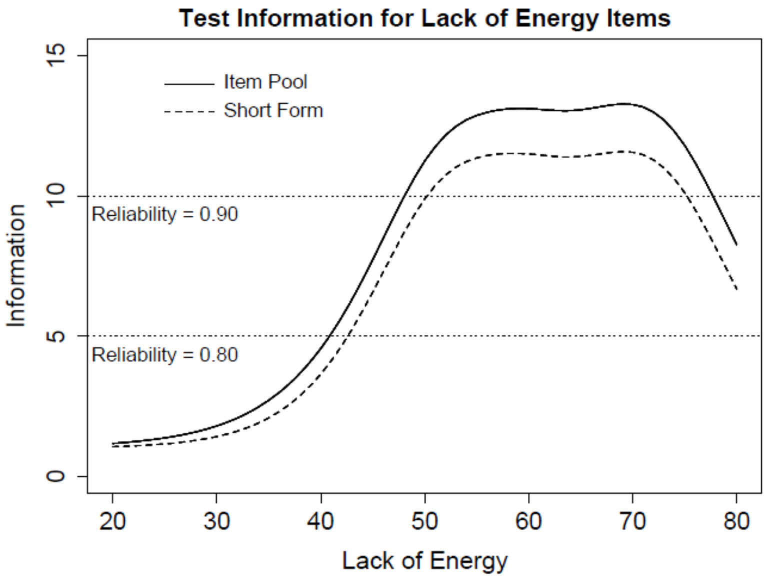

Figure 3.
Test information curves are displayed for the 23 *Tired* item pool, the 8 item short form, and the most informative 8 items at various score location.

Table 1

Study participant characteristics

	Form 1 n=759(%)	Form 2 n=770 (%)	Form 3 n=754 (%)	Form 4 n=765 (%)	Total n= 3,048 (%)
Child’s Gender					
Male	382 (50)	351 (46)	355 (47)	382 (50)	1,470 (48)
Female	377 (50)	419 (54)	399 (53)	383 (50)	1,578 (52)
Child’s Age (yrs)					
8–12	446 (59)	441 (56)[a]	303 (40)	426 (56)	1,616 (53)
13–17	312 (41)	326 (42)[a]	451 (60)	337 (44)	1,426 (47)
Child’s Race					
White	457 (60)	452 (59)	457 (61)	462 (60)	1,828 (60.0)
Black or African-American	154 (20)	168 (22)	172 (23)	150 (20)	644 (21.1)
Others (including multi-races)	122 (16)	128 (17)	100 (13)	129 (17)	479 (16)
Missing	26 (3)	22 (3)	25 (3)	24 (3)	97 (3)
Child’s Ethnicity					
Non Hispanic	614 (81)	641 (83)	617 (82)	619 (81)	2,491 (82)
Hispanic	141 (19)	121 (16)	131 (17)	141 (18)	534 (18)
Missing	4 (1)	8 (1)	6 (1)	5 (1)	23 (1)
Child’s Chronic Conditions Past 6 mo					
No	600 (79)	580 (75)	569 (76)	592 (77)	2,341 (77)
Yes	157 (21)	187 (24)	180 (24)	169 (22)	693 (23)
Missing	2 (0)	3 (0)	5 (1)	4 (1)	14 (1)
Guardian’s* Education Level					
Less than high school	51 (7)	50 (7)	67 (9)	71 (9)	239 (8)
High school graduate/GED	151 (20)	153 (20)	163 (22)	159 (21)	626 (21)
Some college/technical degree	255 (34)	245 (32)	251 (33)	260 (34)	1011 (33)
College degree	179 (24)	214 (28)	183 (24)	180 (24)	756 (25)
Advanced degree	121 (16)	105 (14)	86 (11)	95 (12)	407 (13)
Missing	2 (3)	3 (0)	4 (1)	0	9 (0)
Data Collection Site					
Schools – NC	57 (9)	57 (7)	49 (7)	51 (7)	214 (7)

	Form 1 n=759(%)	Form 2 n=770 (%)	Form 3 n=754 (%)	Form 4 n=765 (%)	Total n= 3,048 (%)
Clinics (NC, TX)	349 (46)	350 (46)	343 (46)	351 (46)	1,393 (46)
Clinics – TX	353 (47)	363 (47)	362 (48)	363 (47)	1,441 (47)

[*] guardian, parent or caregiver completing sociodemographic form and signing consent documents

Table 2

Item Descriptions across Forms

Form	Item Stem	Concept		Correlation between Concepts
		Tired	(Lack of) Energy	
1	Being tired made it hard for me to do things that I usually do.	0.74		NA[e]
	I needed to sleep during the day.	0.65		
	I was too tired to read.	0.62		
	I was too tired to eat. [a]	0.63		
	I did not have much energy. [b]	0.69		
	I had enough energy to eat.[a,b]	0.29		
2	I had enough energy to do things outside.		0.83	0.49
	I had enough energy to do the things I like to do.		0.81	
	I felt strong (not weak).		0.74	
	I had enough energy to read.		0.53	
	I felt too tired to spend time with my friends.	0.73		
	Being tired made it hard for me to keep up with my schoolwork.	0.66		
	I had trouble finishing things because I was too tired.	0.62		
	I took a lot of naps.	0.49		
3	I had enough energy to do my usual things at home.		0.76	0.74
	I had enough energy to play or go out with my friends.		0.71	
	I had enough energy to take a bath or shower.		0.60	
	I felt full of energy.[c]	0.39	0.35	
	I had trouble starting things because I was too tired.	0.71		
	I was too tired to do sports or exercise.	0.70		
	I was so tired it was hard for me to pay attention.	0.68		
	Being tired kept me from having fun.	0.68		
	I was too tired to do things outside.	0.66		
	I was too tired to go out with my family.	0.66		
	I felt more tired than usual when I woke up in the morning.	0.64		
	I felt tired.	0.61		
4	I had enough energy to go out with my family.		0.74	0.73
	I had energy.		0.72	
	I had enough energy to do sports or exercise.		0.72	
	I had enough energy to focus on my work. [d]		0.73	
	I was so tired it was hard for me to focus on my work. [d]	0.64		
	I was too tired to enjoy the things I like to do.	0.72		
	Being tired made it hard for me to play or go out with my friends as much as I'd like.	0.69		
	I got tired easily.	0.67		
	I felt weak.	0.67		

Form	Item Stem	Concept		Correlation between Concepts
		Tired	(Lack of) Energy	
	I was too tired to take a bath or shower.	0.56		
	I was too tired to go up and down a lot of stairs.	0.55		
	It was hard for me to get out of bed in the morning because I was too tired.	0.52		
	I was too tired to watch television.	0.44		

[a] The residual correlation between these items was *r* = 0.42.

[b] Not included in the final calibration of the *Tired* item bank.

[c] Calibrated as part of “(Lack of) Energy”

[d] The residual correlation between these items was *r* = 0.39.

[e] No correlation was estimated due to a small number of items included in the (Lack of) Energy dimension.

Table 3

Summary of the Item Parameter Estimations, Fit Statistics and Evaluation of Gender Differential Item Functioning

Table 3a: “*Tired*” *Items*

Short Form (X)	**Item Stem (Items in final item bank)**	**Item Parameters**					**$S\text{-}X^2$ Fit Statistics**	**LR DIF**		
		a	***b1***	***b2***	***b3***	***b4***	***p***	***χ2***	***d.f.***	***p***
X	I was too tired to enjoy the things I like to do.	1.90	−0.05	0.91	1.94	2.62	0.61	9.3	5	0.10
X	Being tired made it hard for me to play or go out with my friends as much as I’d like.	1.82	−0.01	1.00	2.04	2.84	0.23	4.1	5	0.54
X	I had trouble starting things because I was too tired.	1.76	−0.72	0.37	1.7	2.92	0.99	16.1	5	0.01
X	Being tired made it hard for me to keep up with my schoolwork.	1.69	−0.17	0.55	1.56	2.29	0.02	10.7	5	0.06
X	I had trouble finishing things because I was too tired.	1.68	−0.87	0.09	1.30	2.18	0.00	3.1	5	0.68
X	I got tired easily.	1.67	−1.34	−0.15	1.38	2.62	0.39	3.1	5	0.68
X	I felt weak.	1.67	−0.51	0.55	1.86	3.05	0.09	10.5	5	0.06
X	I was so tired it was hard for me to pay attention.	1.64	−0.72	0.28	1.53	2.39	0.56	9.8	5	0.08
	I felt too tired to spend time with my friends.	1.56	0.54	1.37	2.25	2.81	0.32	4.2	5	0.52
X	I was too tired to do sports or exercise.	1.55	−0.20	0.78	1.89	2.58	0.56	14	5	0.02
X	I was too tired to do things outside.	1.53	−0.60	0.44	1.55	2.36	0.83	10.5	5	0.06
	I needed to sleep during the day.	1.48	−0.63	0.06	1.31	2.28	0.50	11.5	5	0.04
	Being tired kept me from having fun.	1.46	0.00	0.85	1.98	2.65	0.47	4.3	5	0.51
	I was so tired it was hard for me to focus on my work.	1.45	−1.14	−0.13	1.92	2.89	0.86	10.9	5	0.05
	I was too tired to go out with my family.	1.42	0.49	1.48	2.27	2.84	0.01	7.8	5	0.17
	I was too tired to eat.	1.39	0.89	1.52	2.67	3.56	0.13	4.3	5	0.51
	I felt more tired than usual when I woke up in the morning.	1.32	−1.22	−0.30	1.13	1.92	0.51	4.3	5	0.51
	I felt tired.	1.31	−2.47	−1.26	1.37	2.91	0.21	7.7	5	0.17
	I was too tired to go up and down a lot of stairs.	1.24	−0.15	0.91	2.06	2.91	0.84	5.1	5	0.40
	I was too tired to read.	1.22	−0.25	0.58	1.94	2.69	0.39	4.9	5	0.43
	I was too tired to take a bath or shower.	1.16	0.46	1.27	2.52	3.18	0.38	8.8	5	0.12
	It was hard for me to get out of bed in the morning because I was too tired.	0.97	−1.64	−0.53	0.99	2.07	0.72	4.6	5	0.47
	I was too tired to watch television.	0.91	0.38	1.41	3.40	4.60	0.89	6.2	5	0.29

Table 3b: “*Energy*” *Items*

Short Form (X)	Item Stem(Items in the final item bank)	Item Parameters					$S\text{-}X^2$ Fit Statistics	LR DIF		
		a	*b1*	*b2*	*b3*	*b4*	*p*	*χ2*	*d.f.*	*P*
X	I had enough energy to do the things I like to do.	2.58	0.23	0.98	1.87	2.30	.71	19.5	5	0.00
X	I had enough energy to do things outside.	2.34	−0.03	0.71	1.65	2.10	.01	11.9	5	0.04
X	I had energy.	2.23	0.20	1.11	2.28	2.54	.24	7.2	5	0.21
X	I felt strong (not weak).	2.17	−0.24	0.56	1.68	2.05	.00	11.5	5	0.04
X	I had enough energy to focus on my work.	1.86	−0.16	0.71	1.87	2.54	.08	7.6	5	0.18
X	I had enough energy to go out with my family.	1.84	0.46	1.27	2.01	2.30	.01	5.6	5	0.35
X	I had enough energy to do sports or exercise.	1.83	−0.04	0.84	1.81	2.40	.04	3.8	5	0.58
X	I had enough energy to play or go out with my friends.	1.83	0.25	1.09	1.84	2.18	.00	27.8	5	0.00
	I felt full of energy.	1.54	−0.62	0.53	2.12	2.96	.00	8.3	5	0.14
	I had enough energy to take a bath or shower.	1.45	1.10	1.91	3.13	3.51	.80	5.1	5	0.40
	I had enough energy to read.	1.09	−0.22	0.72	2.03	2.43	.01	8.2	5	0.15

Notes: Items “I took a lot of naps” and “Being tired made it hard for me to do things that I usually do” showed gender DIF and were set aside.

NOTE: Item “I had enough energy to do my usual things at home” showed gender DIF and was not included in the final calibration.

Table 4

PROMIS Pediatric Fatigue (Lack of Energy and Tired) Scales Summed Score to Scale Score Translation for Recommended Short Forms

Lack of Energy		
Summed Score	**Scaled Score (*T*)**	**Standard Error**
0	36	5.9
1	42	4.5
2	44	4.2
3	46	3.8
4	48	3.7
5	50	3.5
6	51	3.4
7	52	3.3
8	54	3.2
9	55	3.2
10	56	3.2
11	57	3.2
12	58	3.2
13	59	3.1
14	60	3.1
15	61	3.1
16	62	3.1
17	63	3.1
18	64	3.1
19	65	3.1
20	66	3.1
21	67	3.1
22	68	3.1
23	69	3.1
24	70	3.1
25	71	3.1
26	72	3.2
27	73	3.2
28	74	3.4
29	75	3.4
30	77	3.7
31	78	3.7
32	81	4.4

Tired		
Summed Score	**Scaled Score (*T*)**	**Standard Error**
0	30	5.5
1	34	4.7

Tired		
Summed Score	**Scaled Score (*T*)**	**Standard Error**
2	37	4.4
3	39	4.1
4	41	3.9
5	43	3.8
6	44	3.7
7	45	3.6
8	47	3.5
9	48	3.5
10	49	3.4
11	50	3.4
12	51	3.4
13	52	3.4
14	54	3.4
15	55	3.4
16	56	3.4
17	57	3.4
18	58	3.3
19	59	3.3
20	60	3.3
21	61	3.3
22	62	3.3
23	63	3.3
24	64	3.3
25	65	3.3
26	66	3.3
27	67	3.3
28	68	3.3
29	69	3.3
30	70	3.3
31	71	3.3
32	72	3.4
33	73	3.4
34	74	3.4
35	76	3.5
36	77	3.6
37	79	3.7
38	80	3.8
39	82	4.0
40	85	4.5

NOTE: Scale scores are on a *T*-score scale; the values of SD are reported as conditional standard errors of measurement