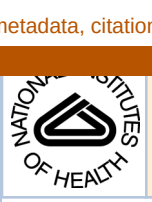

Published in final edited form as:

Pain. 2010 July ; 150(1): 173–182. doi:10.1016/j.pain.2010.04.025.

Development of A Promis Item Bank to Measure Pain Interference

Dagmar Amtmann[1], **Karon F. Cook**[1], **Mark P. Jensen**[1], **Wen-Hung Chen**[2], **Seung Choi**[3], **Dennis Revicki**[2], **David Cella**[3], **Nan Rothrock**[3], **Francis Keefe**[4], and **Leigh Callahan**[5]

[1] Department of Rehabilitation Medicine, University of Washington, Seattle, WA

[2] Center for Health Outcomes Research, United BioSource Corporation, Bethesda, MD

[3] Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL

[4] Departments of Psychiatry & Behavioral Sciences, Anesthesiology, Medicine and Psychology and Neuroscience: Social and Health Sciences, Duke University and Duke University Medical Center, Durham, NC

[5] Departments of Medicine, Orthopaedics, and Social Medicine and Thurston Arthritis Research Center, University of North Carolina, Chapel Hill, NC

Abstract

This paper describes the psychometric properties of the PROMIS Pain Interference (PROMIS-PI) bank. An initial candidate item pool (n=644) was developed and evaluated based on review of existing instruments, interviews with patients, and consultation with pain experts. From this pool, a candidate item bank of 56 items was selected and responses to the items were collected from large community and clinical samples. A total of 14,848 participants responded to all or a subset of candidate items. The responses were calibrated using an item response theory (IRT) model. A final 41-item bank was evaluated with respect to IRT assumptions, model fit, differential item function (DIF), precision, and construct and concurrent validity. Items of the revised bank had good fit to the IRT model (CFI and NNFI/TLI ranged from 0.974 to 0.997), and the data were strongly unidimensional (e.g., ratio of first and second eigenvalue = 35). Nine items exhibited statistically significant DIF. However, adjusting for DIF had little practical impact on score estimates and the items were retained without modifying scoring. Scores provided substantial information across levels of pain; for scores in the T-score range 50-80, the reliability was equivalent to 0.96 to 0.99. Patterns of correlations with other health outcomes supported the construct validity of the item bank. The scores discriminated among persons with different numbers of chronic conditions, disabling conditions, levels of self-reported health, and pain intensity ($p < 0.0001$). The results indicated that the PROMIS-PI items constitute a psychometrically sound bank. Computerized adaptive testing and short forms are available.

Corresponding author: Dagmar Amtmann, Ph.D., Research Assistant Professor, University of Washington, Department of Rehabilitation Medicine, Box 357920, Seattle, WA 98195-7920, (206) 543-4741 V, (206) 685-9224 FAX, dagmara@u.washington.edu.

Conflict of Interest Statement: None of the authors of this paper have any conflicts of interest, financial or otherwise. The item bank was created using NIH funding and it is free and publicly available. Some of the authors have received funding from pharmaceutical companies and/or industry in past, but this funding does not create a conflict of interest with respect to the work covered in this paper.

Keywords

Quality-of-life outcomes; quality-of-life measurement; pain

Introduction

Pain interference (also known as “pain impact”) refers to the degree to which pain limits or interferes with individuals' physical, mental and social activities. This domain is increasingly recognized as important for both understanding patients' experiences and as a key outcome in pain clinical trials [22]. A number of measures of pain interference have been developed including a 9-item scale from the West Haven-Yale Multidimensional Pain Inventory (WHYMPI) [31], a 7-item scale from the Brief Pain Inventory [14,19], a 6-item Pain Impact Questionnaire (PIQ-6) [2], a 3-item scale from the Chronic Pain Grade [57], and the 7-item Pain Disability Index [42]. Available evidence supports the validity and reliability of these scales as measures of pain interference, although each has strengths and weaknesses [15,31,42,57] (see also reviews [22,36]). One weakness of existing instruments is that they are static measures; they require respondents to complete all items, even items that provide no additional information about a person's level of pain interference. Although there are exceptions [2,34], most prior work in the development of pain interference instruments has been limited to Classical Test Theory approaches (CTT). CTT provides the theoretical and mathematical bases for traditional estimates of reliability and validity [1], however it has limited usefulness in evaluating the functioning of individual response options and how precisely items measure across the continuum of pain interference, from little to severe pain interference [26,27].

An alternative approach to the measurement of patient-reported outcomes (PROs) is to develop banks of items that measure the outcome of interest and calibrate responses to these items using an item response theory (IRT) model [17,49,59]. IRT-calibrated instruments provide options such as: (1) computer adaptive testing (CAT), which provides precise measurement using few items; (2) the ability to compute scores that are directly comparable even when respondents take different items, facilitating comparisons across time and between different samples; and (3) development of population-specific short forms containing items of most relevance to a specific population. A priori expectations for what constitutes a psychometrically sound item bank for measuring pain interference include providing reliable scores (Cronbach's alpha > 0.85) with minimal respondent burden, and having a minimal number of items with differential item functioning (DIF). In addition, the validity of the scores should be evidenced by correlations >0.80 with scores on established instruments that measure pain interference.

The National Institutes of Health's (NIH) Patient-Reported Outcomes Measurement Information System (PROMIS) initiative [10,11,20,43,44,50] targeted four pain subdomains: interference, quality, behaviors and intensity. The purpose of this paper is to describe the development of the PROMIS item bank for measuring pain interference. The specific aims of this project were to:

1. Administer candidate items to large community and clinical samples
2. Conduct psychometric analyses, including:
    a. Checking that assumptions for IRT models are met
    b. Fitting a graded response model to the data, examining item fit, removing items that do not fit the model and calibrating final items;

c. Examining evidence for reliability and validity of the PROMIS pain interference bank score

d. Examining DIF

Methods

Development of a Candidate Item Bank

Items that become part of an IRT-calibrated item bank must survive intense psychometric scrutiny; therefore, it is important to test more items than are needed in a final bank. It also is critical that the items adequately sample the content of the domain being measured and, collectively, target all levels of the domain that might be observed in a study population. Development of the candidate PROMIS item bank began with the collection and classification of a “library” of pain interference items (n=644). These items were identified based on an extensive literature review and feedback from persons experiencing pain. Details of the general item development process used by PROMIS have been published elsewhere [20]. Briefly, all pain interference items were reviewed and revised by members of an expert panel of researchers who had expertise in pain assessment, language translation, literacy or psychometrics. After initial revisions were made, persons who had pain participated in in-person interviews to evaluate item clarity, content, and appropriateness. Based on the results of these interviews, item revisions were made. A subset of 56 items was identified to constitute a candidate item bank for measuring pain interference-- the PROMIS-PI item bank. The temporal context for all items was seven days (e.g., “In the past seven days, how much did pain interfere with your enjoyment of life?”). Response options were initially limited to four sets: (1) not at all, a little bit, somewhat, quite a bit, very much; (2) never, rarely, sometimes, often, always; (3) never, once a week or less, once every few days, once a day, every few hours; and (4) never, 1 night, 2-3 nights, 4-5 nights, 6-7 nights. These response categories were examined through in-person interviews to ensure that they were meaningful to and easily understood by individuals living with pain. Participants were asked whether each item had just enough, too few, or too many response options, and whether the response options made sense.

Collection of Item Responses from Clinical and Community Samples

Two types of sources were used to recruit subjects for the current analyses: (1) the PROMIS Wave I data collection and (2) two additional community samples of individuals who were likely to experience chronic pain. The PROMIS Wave 1 data collection served the diverse psychometric needs of developing 14 different item banks. To meet these needs a large and complex sampling design was developed. Detailed description of the PROMIS sampling design is outside the scope of this paper, however, a diagram and description of the Wave I data collection is presented in detail on the PROMIS website (www.nihpromis.org) under the subheading “Wave 1 Testing”.

PROMIS Wave I

PROMIS Metric: With IRT models, scores representing the domain of interest are calibrated in logits that generally range from approximately -4 to +4 [26]. Because it is not intuitive to report health outcomes in negative numbers, researchers typically do a linear transformation of IRT-calibrated scores so that all scores are positive. By consensus, PROMIS decided that the metric for all PROMIS measures would be the T-score metric [35] in which scores have a mean of 50 and a standard deviation (SD) of 10. The advantage of this scoring approach is that a person's score on any PROMIS measure communicates that person's level of the domain relative to the general population. For example, a person who

has a PROMIS-PI score of 70 has a pain interference level approximately two SDs above the estimated national average.

Wave I Measures: In a large-scale data collection (Wave 1 testing), responses were collected to items of 14 PROMIS candidate item banks including the Pain Interference bank. The process described above in which a "library" of 644 pain interference items was reduced to a candidate item bank of 56 was paralleled for all 14 PROMIS domains. Responses to a total of 784 PROMIS items representing the 14 domains were tested in Wave 1. For validity testing, items from "legacy measures" also were administered. A legacy measure is defined as one that has demonstrated reliability and validity and is widely accepted in the relevant application. The legacy measures for testing the validity of the PROMIS-Pain Interference (PI) bank included the 7-item pain interference subscale of the Brief Pain Inventory (BPI) [13] and the 2-item Bodily Pain Subscale of the Medical Outcomes Short Form-36, version 2-acute [60,61].

In addition to PROMIS candidate items and legacy items, two global health items were administered. Respondents were asked, "In general, would you say your health is excellent/ very good/good/fair/poor," and, "How would you rate your pain on average?" ("0=no pain" through "10=worst imaginable pain").

Wave I Sample: A total of 21,133 total persons participated in Wave 1 testing. Of these, 1,532 were recruited from PROMIS research sites. The other 19,601 were recruited through YouGovPolimetrix (www.polimetrix.com, also see www.pollingpoint.com), a polling firm based in Palo Alto, California. Data obtained through YouGovPolimetrix and PROMIS research sites were collected using websites on secure servers. The primary research sites and the YouGovPolimetrix sample included both community and clinical samples. The clinical samples included persons with heart disease (n = 1,156), cancer (n = 1,754), rheumatoid arthritis (n = 557), osteoarthritis (n = 918), psychiatric illness (n = 1,193), COPD (n = 1,214), spinal cord injury (n = 531), and other conditions (n = 560).

The large number of items made it impractical to administer all items to all persons. To obtain sufficient responses to the items of the 14 candidate PROMIS item banks, data collection was divided into two arms. In a "full bank" testing arm, 7,005 persons were administered all items of two of the fourteen, 56-item, candidate PROMIS item banks. For example, one group of participants responded to all 56 items measuring pain interference and all 56 items measuring pain quality. Another group responded to all 56 items of the PROMIS anger item bank and all 56 items of the depression item bank. The second arm of Wave I data collection was a "block testing" arm. For this data collection arm, the 56 items of each item bank were divided into eight, 7-item blocks (8 blocks × 7 items = 56 items of the item bank).

Respondents were included in the Wave I analysis if they responded to 50% or more of the items, did not have repetitive strings of ten or more identical responses (except for response strings of "no pain"), and their response time was greater than one second per item. A total of 14,848 persons met inclusion criteria and responded either to the full, 56-item PROMIS-PI bank (N=845) or to one of the eight, 7-item Pain Interference blocks (N=14,003).

Supplementary Data Collection—In spite of the large data sample in the Wave I data collection, there were relatively few persons who reported severe levels of pain (that is, pain intensity scores of 7 or greater on a 0-10 scale). For 47 of the 56 candidate PROMIS-PI items, more than half of respondents indicated no pain interference (range across items = 27.7% to 79.3% reporting no pain interference). At the other end of the pain interference continuum, for 49 of the 56 items, the highest category (indicating highest level of pain

interference) was endorsed by 5% or less of the sample (range across items = 0.6% to 25.7%). In order to obtain accurate item parameter estimates for items that target higher levels of pain intensity, we added data collected from participants with cancer as part of a project conducted earlier by the PROMIS Statistical Coordinating Center, and we initiated an online survey through the American Chronic Pain Association (ACPA). Before starting the ACPA data collection, nine items were removed from the candidate PROMIS-PI bank based on initial psychometric analyses following Wave I data collection and secondary review by content experts. Of the nine items removed, five were removed because of poor fit, three items were removed because they did not specifically mention pain, and one item was removed because of poor correlation with other items in the bank. This left a revised candidate item bank with 47 items.

Cancer Sample (n=532): 1,754 patients with cancer were included in the PROMIS Wave I sample. Because of the high incidence of pain in persons with cancer [8], we anticipated that including additional patients with cancer likely would increase endorsement of response categories indicating higher pain interference. To obtain these additional participants, we used data that the PROMIS Statistical Coordinating Center had collected from a convenience sample of 532 persons with cancer. This study included persons with any cancer diagnosis at any stage, severity, or treatment status (active or follow-up). Data were collected from two cancer clinics (NorthShore University HealthSystem and John H. Stroger, Jr. Hospital of Cook County), and from cancer support societies in the Chicago area and across the country. Participants were administered 39 of the 47 items in the revised, candidate Pain Interference bank. This 39-item subset of the 56 candidate PROMIS-PI items was selected based on the needs of the prior study. The BPI Interference scale, the 0-10 numerical rating scale of pain intensity items (present, least, worst, and average pain), and SF-36 Bodily Pain Subscale were not administered in this sample.

American Chronic Pain Association (ACPA) Sample (n=523): To increase the clinical sample size further, participants with chronic pain were recruited through the ACPA. An invitation to complete the PROMIS pain survey was posted on the ACPA website. To be eligible, participants had to be 21 years of age or older and have at least one chronic pain condition for at least 3 months prior to participating in the survey. Those who met eligibility criteria were asked to provide informed consent. Those who did so were immediately administered the survey that consisted of the 47-item PROMIS Pain Interference candidate bank as well as clinical items and demographic questions. Respondents were not paid for their participation, reducing the likelihood that individuals without chronic pain would fill out the survey in order to receive a stipend. PROMIS assessment center was used for the data collection and data were screened for quality using the same procedures used for all PROMIS data (e.g., time it took to complete each question, strings of the same responses). To limit response burden, the BPI Interference Subscale, numerical rating scale, and SF-36 Bodily Pain Subscale were not administered in this sample. The invitation to participate in the survey was posted on the ACPA website from August 2007 to February 2008.

Responses to Items Related to Emotional Aspects of Pain: Different types of chronic pain may be associated with different responses to items related to emotional aspects of pain. To ensure that the Wave I, cancer, and chronic pain samples did not differ significantly in their emotional responses to pain, we calculated the mean item responses (range of 1-5) to the 7 items in the bank that referenced emotional components of pain (e.g., “How often did pain make you feel depressed?”). We compared these mean item scores to mean scores on the non-emotional items for participants in each of our 3 samples—those with chronic pain, those with cancer, and those from the PROMIS Wave I sample. Table 1 reports the results. We found that patterns of responses of those with cancer were very similar to those in the Wave I sample, both with respect to mean item scores and mean differences in item scores.

Compared to those with cancer and those in the Wave 1 sample, those with chronic pain endorsed higher response categories on average for both emotional and non-emotional pain impact items. Differences between emotional and non-emotional items were also higher in this group than for the other two samples. However, the differences in all groups were relatively small—less than one quarter of a category on average for the chronic pain sample (on a scale from 1 to 5).

Reduction of Candidate Item Pool to 41-Item Bank—The combined data from the Wave I sample and from the auxiliary samples (cancer and chronic pain patients) were used to conduct analyses of the 47-item candidate item pool. Based on evaluations of dimensionality, fit to the Graded Response Model (GRM) [51], and additional review by content experts, the number of items was reduced to 41 (see Appendix A for a list of the final 41 item bank of Pain Interference items). To evaluate the content representation of the final bank, we compared the content subdomains represented in the original 56 items to those represented in the final 41-item bank. Eleven hypothesized pain interference subdomains were represented in the original item bank; items targeting ten of these were retained in the final bank. Because of concerns about multidimensionality, poor model fit, and/or discriminant validity, the single item related to sex was dropped as were 4 of 5 sleep items, 3 of 4 travel items, and 3 of 4 walking items. The majority of items representing all other pain interference subdomains were retained. In the 41-item bank, 8 items target activities of daily living and work; 4 cognition; 7 emotional function; 4 fun, recreation, and leisure; 1 sleep; 10 social function; and 7 sitting, walking and standing. All analytic results reported below are based on this 41-item bank. Item elimination also reduced the original four response sets to three response sets in the final 41-item bank.

Analyses

IRT Assumptions

IRT is a probability model and estimation of scores is achieved without the requirement that every person respond to the same items [23]. Unidimensional IRT models (as do CTT models) make the assumption that a single latent construct drives the variance in scores. Health outcomes are conceptually complex and responses to items of health outcome measures rarely, if ever, meet a strictly interpreted unidimensionality assumption [18,45,46]. The pertinent question is whether the presence of secondary dimensions disturbs parameter estimates when responses are calibrated using unidimensional IRT models [47,48]. The typical approach to evaluating this question is to assess the strength of a common dimension. The published PROMIS analysis plan [44] suggested the use of confirmatory factor analysis (CFA) in which a unidimensional model is applied and fit statistic values are compared to prior published criteria by Hu and Bentler [3,28,29], McDonald [37], and others [6,33,62]. These criteria included: Comparative Fit Index (CFI) >0.95, Root Mean Square Error of Approximation (RMSEA) <0.06, Tucker Lewis Index or Non-normed Fit Index (NNFI) >0.95, and Standardized Root Mean Square Residual (SRMR) <0.08 [44]. Recognizing the difficulty of attaining CFA standards in the context of item banking (e.g., because of large number of items) alternative standards based on exploratory factor analysis (EFA) results were suggested [44]. As stated in the PROMIS analysis plan, support for unidimensionality is considered sufficient when the first factor accounts for at least 20% of the variability, the ratio of first and second eigenvalues is greater than 4, and the results of scree test, correlations among factors, and factor loadings support the hypothesized structural patterns [44]. In addition, we conducted a parallel analysis of the raw PROMIS data based on principal axis/common factor analysis and on the distribution of the raw data set [39]. The magnitudes of eigenvalues expected by chance alone were computed and compared to observed eigenvalues to estimate the maximum number of underlying dimensions. If the

results of the CFA and EFA analyses produced evidence for secondary dimensions, bi-factor model analyses were planned to assess the level of disturbance in item parameter estimates attributable to multidimensionality.

A second assumption of IRT models is local independence. Local independence holds if, after accounting for the dominant factor, there is no significant association among item responses [52,63]. Local dependency (LD) can adversely impact IRT parameter estimates. To evaluate LD, we examined the residual correlation matrix produced by the single factor CFA. Residual correlations whose absolute values are greater than 0.10 [44] are of some concern, but of particular concern are absolute residual correlations greater than 0.20 [18].

IRT Calibration and Fit

We modeled responses to the 41 items of the candidate PROMIS-PI items using Samejima's two parameter polytomous graded response model (GRM) [51] with Multilog, Version 7.03 [55]. GRM is an IRT model suitable for ordered polytomous responses. One of the advantages of IRT, as compared to CTT, is that IRT is based on a mathematical model that allows comparison of the patterns of actual responses to those predicted by the model. How closely the actual data correspond to the predictions of the model can be quantified and summarized by goodness-of-fit statistics.

Fit to the GRM was calibrated using the computer macro, IRTFIT [5]. We report values of S-X^2 (a Pearson X^2 statistic) and S-G^2 (a likelihood ratio G^2 statistic) [40,41]. These statistics compare expected and observed frequencies of item category responses for various levels of scores and quantify the differences between expected and observed responses.

Reliability Analysis

In CTT, single reliability estimates are calculated for the entire scale despite the fact that the scale is likely to provide more precise measurement at different levels of the domain being measured. In IRT the concept of reliability is conceptualized as “information” and extended to take into account the fact that measurement precision can differ across levels of the domain being measured. The relationship between standard error (SE) and information is defined by the formula, $SE(\theta)=1/\sqrt{I(\theta)}$, where *SE* is the standard error of the posterior distribution, *I* is information, and θ is estimated domain level (from no or mild pain interference to high levels of pain interference). As the formula indicates, increased scale information is associated with smaller SE's and, therefore, greater precision. To determine the effects of domain level on precision, we plotted information for the 41-item PROMIS-PI item bank and compared the results to the distribution of Pain Interference scores in the calibration samples.

Validity Analyses

The construct validity of the PROMIS-PI was evaluated by comparing PROMIS-PI scores to scores from measures of similar domains [9] including the BPI Interference Subscale [13], the SF-36 (version 2-acute, 7 day recall) Bodily Pain Subscale (SF-36 BP) [60,61], and a 0-10 Numerical Rating of Pain Intensity [30]. Also evaluated was the association between PROMIS-PI scores and scores on PROMIS measures of theoretically different domains (i.e., anxiety, depression, fatigue, and physical function).

To evaluate concurrent validity, we assessed how well PROMIS-PI scores distinguished between and among “known groups.” That is, we evaluated whether PROMIS-PI scores distinguished among subgroups that, theoretically, should differ in mean scores. Analyses of Variance (ANOVAs) were conducted to compare mean scores of those who reported different levels of “average health” (excellent, very good, good, fair, poor). ANOVAs also

were conducted to compare PROMIS-PI scores based on numbers of chronic and disabling conditions reported and based on levels of pain intensity.

Differential Item Functioning

Analysis of differential item functioning (DIF) examines the relationships among item responses, levels of the domain being measured, and subgroup membership. For any given level of domain, the probabilities of endorsing specified item responses should be independent of subgroup membership [25]. In the context of pain interference measurement, persons of different ages, education-levels, race/ethnicities, and genders who have equal levels of pain interference should be equally likely to endorse a particular category of a specified PROMIS-PI item. For example, men and women *who are equal in their levels of pain interference* should be equally likely to respond, "somewhat" to the item, "How much did pain interfere with work around the home?" When items function similarly across demographic groups (do not exhibit DIF), direct comparison of group scores are justified, even when persons respond to different items from the item bank. DIF can be consistent across the range of the domain being measured (uniform DIF), or its impact can vary for persons with different levels of the domain being measured (non-uniform DIF). For the current study, DIF analyses were conducted with freeware developed for testing DIF using ordinal logistic regression (OLR) [12,64]. Likelihood-ratio χ^2 statistics compared OLR models with and without subgroup membership as a predictor. DIF was evaluated with regard to age (under 65 versus 65 and over), education-level (high school or less versus some college or higher), and gender (male versus female).

Results

Sample Characteristics

Table 2 provides demographic details for the PROMIS Wave I data and for the auxiliary samples. Across all samples, the majority of respondents were white (83.4%) and female (54.0%). Respondents were well educated; 44.9% had at least a college degree and 79.7% had at least some college or technical school. The average age of the combined sample was 54 years.

IRT Assumptions

Dimensionality assessment was evaluated using CFA fit statistics, EFA results, and parallel analysis. The chi-square statistic is sensitive to sample size [7] and as expected it was statistically significant (χ^2 (68, N = 15,903) = 2624, $p < .001$). The CFI and NNFI/TLI values were very high at 0.974 and 0.997, respectively, well beyond published fit criteria of >0. 95. Likewise the SRMR value of 0.033 was well below the published criterion of < 0.08. However, the RMSEA was 0.175, which is well beyond the published criterion of < 0.06. This finding is not surprising given our previous work that found that RMSEA values tend to be elevated when there are larger numbers of items, and this statistic may not be appropriate for judging dimensionality of a large item bank [16].

EFA results supported sufficient unidimensionality. The first factor accounted for 86% of the variance; and the second factor accounted for 2% of the variance. The first eigenvalue was 35.3 and the second was 1.0. Moreover, the first and second factor identified by the EFA were highly correlated (r= 0.76). These results concur with those of the parallel analysis; only the first factor had an eigenvalue greater than that expected by chance alone. We examined the items that loaded higher on the second factor than on the first factor to evaluate whether a clinically interesting second factor could be identified and potentially scored. The items with higher loadings on the second factor related to working, doing household chores, avoidance of social activities because of pain, and difficulties with sitting

for a period of time and thus did not provide evidence for a conceptually distinct secondary dimension. Because the combined results of these analyses strongly suggested a single dominant factor, we did not conduct follow-up bi-factor model analysis.

Examination of the residual correlations indicated very little local dependence. The average absolute value of the residual correlations was 0.03. A total of 8 (<1 %) were greater than 0.10, and none were greater than 0.13. The three possible pairs of items PI50, PI51, and PI55 all had residual correlations of 0.13. Each of these items asks about pain interference on sitting. Based on these results we judged the level of local dependency to be minor and not to pose a substantial threat to the accuracy of IRT parameter estimation.

IRT Calibration and Fit

As described above, the metric of the calibration was anchored to a community sample and weighted to reflect the 2000 United States Census [35]. Scores are reported on a T-score metric and ranged from 40.1 to 80.4, roughly one SD below to 3 SD above the population mean. Appendix A presents the content, response sets, and item parameters of the PROMIS-PI. Pain interference items were scored so that higher scores indicated greater pain interference. Figure 1 displays the distribution of scores by sample. Only scores for persons who responded to at least seven items (the number of items administered in the Wave I block design) are reported here. As the plots show, and as mentioned above, a large portion of the PROMIS Wave I respondents reported experiencing no pain interference; 33% answered “not at all” or “never” to every item. Of the respondents from the cancer sample obtained from the PROMIS Statistical Coordinating Center, 21% reported no pain interference. As expected, all of the ACPA respondents reported at least some pain interference. Based on an alpha value of 0.01, no items were found to misfit the GRM. Probability values for S-X^2 statistics ranged from 0.102 to 0.998 (mean = 0.761). Probabilities for S-G^2 statistics ranged from 0.053 to 0.998 (mean = 0.731).

Reliability Analyses

The Cronbach alpha estimate for the PROMIS-PI bank was 0.99. Figure 2 is a plot of total information by Pain Interference T-scores for the combined samples (PROMIS Wave I, ACPA, and Cancer samples). Reference lines indicate levels of information approximately equivalent to reliability estimates of 0.90 and 0.95. As the plot demonstrates, the bank provides substantial information across levels of domain observed in the combined sample (T-score range 40.1-80.4). In fact, for the range of the mean (score of 50) to three standard deviations above the mean (score of 80), the bank provides information equivalent to reliability of 0.96 to 0.99.

Validity Analyses

Table 3 reports correlation coefficients that estimate the associations between PROMIS-PI scores and scores on other PROMIS measures and other pain-related measures. Pearson r was used to correlate PROMIS scores (thetas) because they met assumptions of equal distances between the units of the scale. A non-parametric indicator (Spearman Rho) was used to estimate associations between PROMIS pain interference scores and ordinal-level scores, i.e., the pain intensity score, the SF-36 BP score, and the BPI interference score. The PROMIS-Physical Function items and the SF-36 BP Subscale are scored so that higher scores indicate greater function and less pain, respectively. The pattern of associations is consistent with expectations and supports the construct validity of the PROMIS-PI item bank. As expected, the strongest correlations were with measures of pain-related domains—BPI Interference Subscale (rho = 0.90), SF-36 Bodily Pain Subscale (rho = - 0.84), and 0-10 Numerical Rating of Pain Intensity (rho = 0.48). Also as expected, the associations were

weakest with PROMIS measures of mental health domains (r = 0.33 with PROMIS-Depression, r = 0.35 with PROMIS-Anxiety).

Table 4 reports the results of ANOVAs evaluating the PROMIS-PI scores in discriminating known groups. As indicated in the table, the PROMIS-PI score means increased stepwise with increases in number of chronic conditions, number of disabling conditions, and decreases in reported general health. These comparisons were all statistically significant with probabilities < 0.0001.

Differential Item Function Detection

None of the items had statistically significant education DIF, nor was any statistically significant non-uniform DIF found among the 41 PROMIS-PI items. However, nine items had statistically significant uniform DIF. One item had gender-related DIF—Item P3, “How much did pain interfere with your enjoyment of life?” Eight items had statistically significant age-related DIF—Items P1, P8, P11, P40, P42, P47, P49, and P56 (see Appendix A). Of these, three related to the cognitive interference of pain (taking in information, concentrating, remembering); two referred to emotional interference (emotionally tense, irritable); and three referred to walking or standing. We evaluated the practical impact of DIF by calibrating group-specific item parameters for all items with statistically significant DIF and then comparing scores based on these “corrected” item parameters to those obtained with the original parameters. Accounting for the one item with gender DIF had negligible effects on individual scores. The differences in T-Scores (original minus DIF-free estimates) ranged from -1.79 to 4.44 (observed T-Scores ranged from 40.1 to 80.4). Only 8 participants (0.6%) had score differences greater than ± 2.50 (± 2 times the median standard error). The impact of age-related DIF was more substantial. Of 1,276 respondents, 37 (2.9%) had absolute score differences greater than 2.50.

Discussion

The results of this study indicate that the PROMIS-PI items constitute a psychometrically sound item bank for assessing the negative effects of pain on functioning in the range experienced by the vast majority of people who have pain. This conclusion is supported by findings concerning (1) unidimensionality, (2) item fit to the IRT model, (3) reliability of the PROMIS-PI scores across different levels of pain interference, (4) associations between PROMIS-PI scores and other measures, (5) and the independence of function of the large majority of items with respect to subgroup membership.

There are several advantages of the PROMIS-PI over traditional measures of pain interference. The item bank can be used to develop short forms for particular purposes or samples, or the items can be administered using CAT. Scores on short forms and on CAT are reported in the same metric and are directly comparable. Also, through the PROMIS Assessment Center, pain interference can be measured in context with other domains (e.g. depression, anxiety, physical function, fatigue, social health) and scores can be graphically displayed and compared with respect to national and subgroup norms. A major strength of the PROMIS-PI bank is that scores have inherent meaning; they communicate respondents' levels of a domain relative to the general population. Thus, the PROMIS PI bank advances the measurement of pain interference.

Responses to self-reported items measuring complex constructs are never strictly unidimensional. However, the results of our analyses support the conclusion that the pain interference domain, at least as measured by the PROMIS-PI item bank, is a homogenous construct. These results have conceptual as well as psychometric implications. Psychometrically, the results strongly support the use of one summary score. Conceptually,

they suggest that pain interference is a relatively "narrow band" domain [45]. The PROMIS-PI item responses exhibited high internal consistency and a single factor dominated despite inclusion of items representing multiple hypothesized pain interference subdomains. The results are consistent with EFA results from most (but not all) analyses of other measures of pain interference [14,32] and consistent with the high internal consistency estimates obtained for other pain interference measures [4,31,38,53,56,58,59].

The PROMIS-PI scores proved to be highly reliable in the T-score range of 50 and 80. They were less reliable in score ranges representing no pain interference to mild pain interference (e.g., scores that are >1 SD below the population mean) and in score ranges reflecting very severe pain interference (e.g., scores that are >3 SD above the population mean). However, the range of high reliability for the PROMIS-PI scores corresponds to the range reported by most individuals in our samples (see Figure 2). For instance, among the ACPA clinical sample, no individuals had a score lower than 40 and only 5 individuals (less than 1%) had PROMIS-PI score over 80.

Strong support for the validity of PROMIS-PI scores was observed in the pattern of correlations with other measures and the scores ability to discriminate among individuals with different levels and numbers of chronic conditions, disabling conditions, and general health. Pain intensity and pain interference scores had approximately 25% shared variance (rho = .487), suggesting that the pain interference and pain intensity are related, but distinct domains.

The findings concerning differential item function support the use of the PROMIS-PI items (and the interpretation of the PROMIS-PI scores) across samples that differ in educational level (no differences were found) and gender (difference found in only one item). DIF was of somewhat greater concern with respect to age (8 items were found to have statistically, age-related DIF). However, though DIF reached statistical significance, its practical impact on scores was minor. We judged this impact to be negligible, retained all items in the bank, and did not construct scoring tables that would account for these differences. Nevertheless, future studies should evaluate the impact of DIF on PROMIS-PI scores obtained using CAT. Additionally, in selecting items from the bank for short forms, we recommend giving preference to items that exhibited no statistically significant DIF.

Study Limitations and Future Research Directions

Although the findings provide preliminary support for the validity and reliability of the PROMIS-PI item bank, validation is an ongoing process, and no single study can provide all the information needed to fully understand the strengths and weaknesses of a measure. Data for the current study were obtained from large community samples that included healthy individuals as well as individuals with a number of specific health problems. Although these samples represent a large range of individuals, it would be useful to expand the evaluation of the PROMIS-PI to additional subgroups, such as persons with specific pain conditions (e.g., headache, low back pain, post-herpetic neuralgia), individuals with other health conditions who often have pain as a secondary complaint or symptom (e.g., patients with multiple sclerosis, cerebral palsy, neuromuscular disease), in ethnic/racial minority samples and in persons with lower levels of education.

In addition, although we performed a variety of analyses useful for determining the psychometric properties of the PROMIS-PI items, additional analyses would be helpful for interpreting the PROMIS PI scores. The interpretability of the PROMIS metric could be extended by estimating score differences representing meaningful category intervals (e.g., "clinically important difference", "clinically meaningful change"). Another important step, and one seldom undertaken with health outcome measures, is to develop supportable

inferences based on PROMIS-PI scores [52]. These would include, for example, associating PROMIS-PI scores and score changes with clinical markers or “actionable” events, such as change in medication or referral to specialists.

Acknowledgments

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 are Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Laura Lee Johnson, Ph.D. 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.

Appendix A: Item Content, Responses and Parameter Estimates

Item^	Response Set*	Item Stem	b 1	b 2	b 3	b 4	a
PI1	A	How difficult was it for you to take in new information because of pain?	0.784	1.356	1.985	2.617	3.275
PI3	A	How much did pain interfere with your enjoyment of life?	0.059	0.798	1.3	1.794	5.316
PI5	A	How much did pain interfere with your ability to participate in leisure activities?	0.165	0.81	1.29	1.821	5.549
PI6	A	How much did pain interfere with your close personal relationships?	0.571	1.093	1.614	2.1	4.389
PI8	A	How much did pain interfere with your ability to concentrate?	0.338	1.07	1.668	2.329	3.725
PI9	A	How much did pain interfere with your day to day activities?	0.105	0.833	1.362	1.962	6.511
PI10	A	How much did pain interfere with your enjoyment of recreational activities?	0.077	0.706	1.139	1.722	5.718
PI11	A	How often did you feel emotionally tense because of your pain?	0.302	1	1.49	2.16	3.666
PI12	A	How much did pain interfere with the things you usually do for fun?	0.136	0.753	1.181	1.759	5.814
PI13	A	How much did pain interfere with your family life?	0.405	0.973	1.497	2.049	5.642
PI14	A	How much did pain interfere with doing your tasks away from home (e.g., getting groceries, running errands)?	0.374	0.932	1.342	1.894	5.201
PI16	B	How often did pain make you feel depressed?	0.398	1.001	1.756	2.421	3.158
PI17	A	How much did pain interfere with your relationships with other people?	0.517	1.134	1.703	2.45	4.787
PI18	A	How much did pain interfere with your ability to work (include work at home)?	0.193	0.842	1.36	1.875	4.817
PI19	A	How much did pain make it difficult to fall asleep?	0.19	0.911	1.448	2.078	2.911
PI20	A	How much did pain feel like a burden to you?	0.062	0.717	1.166	1.706	4.435

Item^	Response Set*	Item Stem	b 1	b 2	b 3	b 4	a
PI22	A	How much did pain interfere with work around the home?	0.107	0.764	1.249	1.863	5.541
PI24	B	How often was pain distressing to you?	-0.049	0.556	1.281	2.058	3.741
PI26	B	How often did pain keep you from socializing with others?	0.528	1.06	1.632	2.338	4.803
PI29	B	How often was your pain so severe you could think of nothing else?	0.548	1.087	1.802	2.823	3.378
PI31	A	How much did pain interfere with your ability to participate in social activities?	0.405	0.925	1.393	1.959	6.185
PI32	B	How often did pain make you feel discouraged?	0.135	0.698	1.401	2.176	3.596
PI34	A	How much did pain interfere with your household chores?	0.11	0.771	1.269	1.87	5.018
PI35	A	How much did pain interfere with your ability to make trips from home that kept you gone for more than 2 hours?	0.676	1.095	1.523	1.984	4.656
PI36	A	How much did pain interfere with your enjoyment of social activities?	0.275	0.868	1.365	1.909	5.604
PI37	B	How often did pain make you feel anxious?	0.349	0.987	1.715	2.471	3.032
PI38	B	How often did you avoid social activities because it might make you hurt more?	0.483	0.895	1.458	2.166	4.761
PI39	B	How often did pain make simple tasks hard to complete?	0.067	0.655	1.39	2.194	4.073
PI40	B	How often did pain prevent you from walking more than 1 mile?	0.286	0.624	0.977	1.366	3.395
PI42	B	How often did pain prevent you from standing for more than one hour?	0.325	0.7	1.084	1.523	3.155
PI46	B	How often did pain make it difficult for you to plan social activities?	0.398	0.872	1.458	2.02	4.797
PI47	B	How often did pain prevent you from standing for more than 30 minutes?	0.286	0.697	1.153	1.641	3.401
PI48	A	How much did pain interfere with your ability to do household chores?	0.166	0.742	1.243	1.832	4.888
PI49	A	How much did pain interfere with your ability to remember things?	0.901	1.452	1.999	2.546	3.075
PI50	B	How often did pain prevent you from sitting for more than 30 minutes?	0.655	1.15	1.696	2.297	3.206
PI51	B	How often did pain prevent you from sitting for more than 10 minutes?	0.954	1.565	2.207	2.799	2.971
PI52	B	How often was it hard to plan social activities because you didn't know if you would be in pain?	0.631	1.048	1.552	1.984	4.611
PI53	B	How often did pain restrict your social life to your home?	0.446	0.953	1.547	2.264	3.881
PI54	C	How often did pain keep you from getting into a standing position?	0.948	1.401	1.801	2.103	2.524
PI55	B	How often did pain prevent you from sitting for more than one hour?	0.657	1.111	1.675	2.347	2.934
PI56	A	How irritable did you feel because of pain?	0.005	0.884	1.558	2.168	3.035

* For all items, the time frame was the past 7 days. The response sets were:
A = Not at all/A little bit/Somewhat/Quite a bit/Very much

B = Never/Rarely/Sometimes/Often/Always

C = Never/Once a week or less/Once every few days/Once a day/Every few hours

b1, b2, b3, b4 = estimated item category difficulties

a = estimated item discrimination

^ Item names are based on their location in the original candidate item pool; therefore, PI1, for example is the first Pain Interference item.

References

1. Anastasi, A. Psychological Testing. New York: Macmillan Publishing Company; 1988.
2. Becker, J.; Schwartz, C.; Saris-Baglama, RN.; Kosinski, M.; Bjorner, JB. Using item response theory (IRT) for developing and evaluating the Pain Impact Questionnaire (PIQ-6TM); Pain Med. 2007. p. S129-S144.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2662709
3. Bentler P. Comparative fit indices in structural models. Psycho Bull 1990;107:238–46.
4. Bernstein, IH.; Jaremko, ME.; Hinkley, BS. On the utility of the West Haven-Yale Multidimensional Pain Inventory; Spine. 1995. p. 956-63.http://www.ncbi.nlm.nih.gov/pubmed/7644962
5. Bjorner, JB.; Smith, KJ.; Stone, C.; Sun, X. Lincoln, RI: QualityMetric; 2007. IRTFIT: A Macro for Item Fit and Local Dependence Tests under IRT Models. http://outcomes.cancer.gov/areas/measurement/irt_model_fit.html
6. Browne, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, KA.; Long, JS., editors. Testing Structural Equation Models. Newbury Park, CA: Sage Publications; 1993.
7. Bull, CR.; Bull, RM.; Rastin, BC. On the sensitivity of the chi-square test and its consequences; Meas Sci Technol. 1992. p. 789-795.http://www.iop.org/EJ/abstract/0957-0233/3/9/001/
8. Burton, AW.; Fanciullo, GJ.; Beasley, RD.; Fisch, MJ. Chronic pain in the cancer survivor: A new frontier. 2002 NIH State of the Science Statement on Symptom Management in Cancer: Pain, Depression, and Fatigue; NIH Con State of the Sci Statm 1–29. Pain Med. 2007. p. 189-98.http://www.ncbi.nlm.nih.gov/pubmed/17305690
9. Campell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 1959;56:81–105. [PubMed: 13634291]
10. Cella, D.; Gershon, R.; Lai, JS.; Choi, S. The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment; Qual Life Res. 2007. p. 133-41.http://www.ncbi.nlm.nih.gov/pubmed/17401637
11. 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; Med Care. 2007. p. S3-S11.http://www.ncbi.nlm.nih.gov/pubmed/17443116
12. Choi, SW.; Gibbons, LE.; Crane, PK. Development of freeware for an iterative hybrid ordinal logistic regression/IRT DIF. Patient Reported Outcomes Measurement Information System (PROMIS) Conference; Bethesda, MD. 2008.
13. Cleeland, CS. Measurement of Pain by Subjective Report. In: Chapman, CR., editor. Advances in Pain Research and Management. New York, New York: Raven Press; 1989. p. 391-403.
14. Cleeland CS, Ryan KM. Pain assessment: Global use of the Brief Pain Inventory. Ann Acad Med 1994;23(2):129–38.
15. The R Project for Statistical Computing, ‘R Vs. 2.7.2’. 2008. http://www.r-project.org/
16. Cook, KF.; Kallen, MA.; Amtmann, D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption; Qual Life Res. May. 2009 p. 447-460.http://www.ncbi.nlm.nih.gov/pubmed/19294529
17. Cook, KF.; Roddey, TS.; O'Malley, KJ.; Gartsman, GM. Development of a flexilevel scale for use with computer-adaptive testing for assessing shoulder function; J Shoulder Elbow Surg. 2005. p. 90S-94S.http://www.ncbi.nlm.nih.gov/pubmed/15726093
18. Cook, KF.; Teal, CR.; Bjorner, JB.; Cella, D.; Chang, CH.; Crane, PK.; Gibbons, LE.; Hays, RD.; McHorney, CA.; Ocepek-Welikson, K.; Raczek, AE.; Teresi, JA.; Reeve, BB. IRT Health

Outcomes Data Analysis Project: An overview and summary; Qual Life Res. 2007. p. 16p. 121-32.http://www.ncbi.nlm.nih.gov/pubmed/17351824

19. Daut RL, Cleeland CS, Flanery RC. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain 1983;17:197–210. [PubMed: 6646795]
20. DeWalt, DA.; Rothrock, N.; Yount, S.; Stone, AA. Evaluation of item candidates: The PROMIS qualitative item review; Med Care. 2007. p. S12-21.http://www.ncbi.nlm.nih.gov/pubmed/17443114
21. Drasgow F, Parsons CK. Application of unidimensional item response theory models to multidimensional data. Appl Psychol Meas 1983;7:189–99.
22. Dworkin, RH.; Turk, DC.; Farrar, JT.; Haythornthwaite, JA.; Jensen, MP.; Katz, NP.; Kerns, RD.; Stucki, G.; Allen, RR.; Bellamy, N.; Carr, DB.; Chandler, J.; Cowan, P.; Dionne, R.; Galer, BS.; Hertz, S.; Jadad, AR.; Kramer, LD.; Manning, DC.; Martin, S.; McCormick, CG.; McDermott, MP.; McGrath, P.; Quessy, S.; Rappaport, BA.; Robbins, W.; Robinson, JP.; Rothman, M.; Royal, MA.; Simon, L.; Stauffer, JW.; Stein, W.; Tollett, J.; Wernicke, J.; Witter, J. Core outcome measures for chronic pain clinical trials: IMMPACT Recommendations; Pain. 2005. p. 9-19.http://www.ncbi.nlm.nih.gov/pubmed/15621359
23. Embretson, SE.; Reise, SP. Item Response Theory for Psychologists. Mahway, NJ: Lawrence Erlbaum Associates; 2000.
24. Gibbons, LE. Boston, MA: Boston College Department of Economics, Statistical Software Components; 2006. DIFWITHPAR: Stata Module for Detection of and Adjustment for Differential Item Functioning (DIF). http://ideas.repec.org/c/boc/bocode/s456722.html
25. Groenvold, M.; Bjorner, JB.; Klee, MC.; Kreiner, S. Test for item bias in a quality of life questionnaire; J Clin Epidemiol. 1995. p. 48-805.www.ncbi.nlm.nih.gov/pubmed/7769411
26. Hambleton, R.; Swaminathan, H.; Rogers, HJ. Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publishing, Inc.; 1991.
27. Hays, RD.; Bjorner, JB.; Revicki, DA.; Spritzer, KL.; Cella, D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res. 2009. Epub ahead of print http://www.ncbi.nlm.nih.gov/pubmed/19543809
28. Hu, LT.; Bentler, PM. Evaluating model fit. In: Hoyle, RH., editor. Structural Equation Modeling: Concepts, Issues and Applications. Thousand Oaks, CA: Sage Publications; 1995. p. 76-79.
29. Hu LT, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling 1999;6:1–55.
30. Jensen, MP.; Ormel, J.; Keefe, FJ.; Dworkin, SF. Measurement of pain. In: Loeser, JD.; Turk, DC.; Chapman, CR.; Butler, S., editors. Bonica's Management of Pain Media. PA: William & Wilkins; in press
31. Kerns RD, Turk DC, Rudy TE. The West Haven-Yale Multidimensional Pain Inventory (WHYMPI). Pain 1985;23:345–56. [PubMed: 4088697]
32. Klepstad, P.; Loge, GH.; Borchgrevink, PC.; Mendoza, TR.; Cleeland, CS.; Kaasa, S. The Norwegian Brief Pain Inventory Questionnaire: Translation and validation in cancer pain patients; J Pain Symptom Manage. 2002. p. 517-25.http://www.ncbi.nlm.nih.gov/pubmed/12547051
33. Kline, RB. Principles and Practice of Structural Equation Modeling. New York, NY: The Guilford Press; 1998.
34. Lai JS, Dineen K, Cella D, von Roenn J. An item response theory based pain item bank can enhance measurement precision. J Pain Symptom Manage 2005;30:278–88. [PubMed: 16183012]
35. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, Hays RD. Representativeness of the PROMIS Internet Panel. J Clin Epidemiol. In press.
36. McCall, WA. How to Measure in Education. New York: Macmillan; 1922.
37. McDonald, RP. Test Theory: A Unified Treatment. Mahway, NJ: Lawrence Earlbaum; 1999.
38. Mendoza, TR.; Chen, C.; Brugger, A.; Hubbard, R.; Snabes, M.; Palmer, SN.; Zhang, Q.; Cleeland, CS. The utility and validity of the modified Brief Pain Inventory in a multiple-dose postoperative analgesic trial; Clin J Pain. 2004. p. 357-62.http://www.ncbi.nlm.nih.gov/pubmed/15322443

39. O'Connor, BP. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's Map Test; Behav Res Meth Ins C. 2000. p. 396-402.http://www.ncbi.nlm.nih.gov/pubmed/11029811
40. Orlando M, Thissen D. Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. Appl Psychol Meas 2003;27:289–98.
41. Orland M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Appl Psychol Meas 2000;24:50–64.
42. Pollard CA. Preliminary validity study of the pain disability index. Perceptual and Motor Skills 1984;59:974. [PubMed: 6240632]
43. Reeve, BB.; Burke, LB.; Chiang, YP.; Clauser, SP.; Colpe, LJ.; Elias, JW.; Fleishman, J.; Hohmann, AA.; Johnson-Taylor, WW.; Lawrence, W.; Moy, CS.; Quatrano, LA.; Riley, WT.; Smothers, BA.; Werner, EM. Enhancing measurement in health outcomes research supported by agencies within the US Department of Health and Human Services; Qual Life Res. 2007. p. 16p. 175-86.http://www.ncbi.nlm.nih.gov/pubmed/17530449
44. 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); Med Care. 2007. p. S22-31.http://www.ncbi.nlm.nih.gov/pubmed/17443115
45. Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. Psychol Assess 2000;12:287–97. [PubMed: 11021152]
46. Reise, SP.; Haviland, MG. Item response theory and the measurement of clinical change; J Pers Assess. 2005. p. 228-38.http://www.ncbi.nlm.nih.gov/pubmed/15907159
47. Reise, SP.; Morizot, J.; Hays, RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures; Qual Life Res. 2007. p. 19-31.http://www.ncbi.nlm.nih.gov/pubmed/17479357
48. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychol Bull 1993;114:552. [PubMed: 8272470]
49. Revicki, DA.; Cella, DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing; Qual Life Res. 1997. p. 595http://www.ncbi.nlm.nih.gov/pubmed/9330558
50. Rose, M.; Bjorner, JB.; Becker, J.; Fries, JF.; Ware, JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS); J Clin Epidemiol. 2008. p. 17-33.http://www.ncbi.nlm.nih.gov/pubmed/18083459
51. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement 1969;17
52. Steinberg L, Thissen D. Uses of item response theory and the testlet concept in the measurement of psychopathology. Psychol Methods 1996;1:81–97.
53. Tan, G.; Jensen, MP.; Thornby, JI.; Shanti, BF. Validation of the Brief Pain Inventory for chronic nonmalignant pain; J Pain. 2004. p. 133-7.http://www.ncbi.nlm.nih.gov/pubmed/15042521
54. Testa, MA. Interpretation of quality-of-life outcomes: Issues that affect magnitude and meaning; Med Care. 2000. p. II166-74.http://www.ncbi.nlm.nih.gov/pubmed/10982103
55. Thissen, D.; Chen, WH.; Bock, RD. MULTILOG (Version 7). Lincolnwood, IL: Scientific Software International; 2003.
56. Uki, J.; Mendoza, T.; Cleeland, CS.; Nakamura, Y.; Takeda, F. A brief cancer pain assessment tool in Japanese: The utility of the Japanese Brief Pain Inventory--BPI-J; J Pain Symptom Manage. 1998. p. 364-73.http://www.ncbi.nlm.nih.gov/pubmed/9879161
57. Von Korff, M.; Ormel, J.; Keefe, FJ.; Dworkin, SF. Grading the severity of chronic pain; Pain. 1992. p. 133-49.http://www.ncbi.nlm.nih.gov/pubmed/1408309
58. Wang, XS.; Mendoza, TR.; Gao, SZ.; Cleeland, CS. The Chinese version of the Brief Pain Inventory (BPI-C): Its development and use in a study of cancer pain; Pain. 1996. p. 407-16.http://www.ncbi.nlm.nih.gov/pubmed/8951936

59. Ware, JE, Jr. Conceptualization and Measurement of health-related quality of life: Comments on an evolving field; Arch Phys Med Rehabil. 2003. p. S43-51.http://www.ncbi.nlm.nih.gov/pubmed/12692771

60. Ware, JE, Jr. SF-36 Health Survey update; Spine. 2000. p. 3130-9.http://www.ncbi.nlm.nih.gov/pubmed/11124729

61. Ware, JE., Jr; Sherbourne, CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual Framework and Item Selection; Med Care. 1992. p. 473http://www.ncbi.nlm.nih.gov/pubmed/1593914

62. West, SG.; Finch, JF.; Curran, PJ. Issues and Applications Structural Equation Modeling: Concepts. Thousand Oaks, CA: Sage Publications; 1995. Sem with nonnormal variables; p. 56-75.

63. Yen WM. Scaling performance assessments: Strategies for managing local item dependence. J Educ Meas 1993;30:187–213.

64. 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, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.

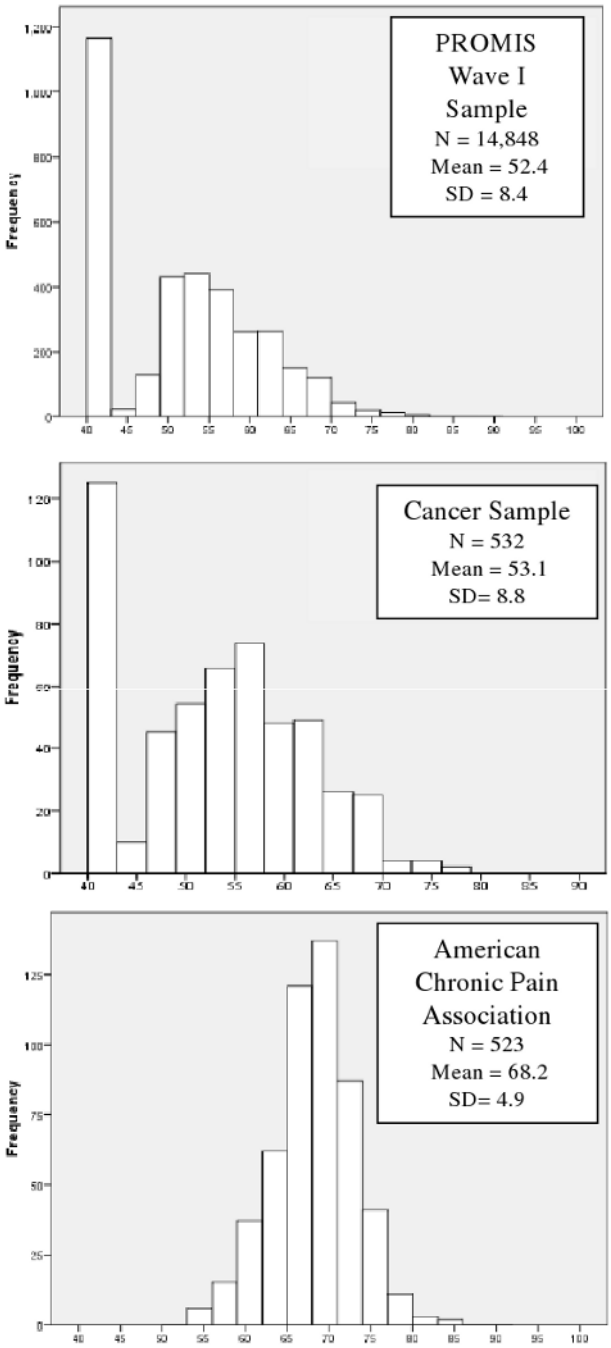

Figure 1. Distribution of PROMIS-Pain Interference T-Scores by Sample (T-Scores Have a Mean of 50 and a Standard Deviation (SD) of 10)

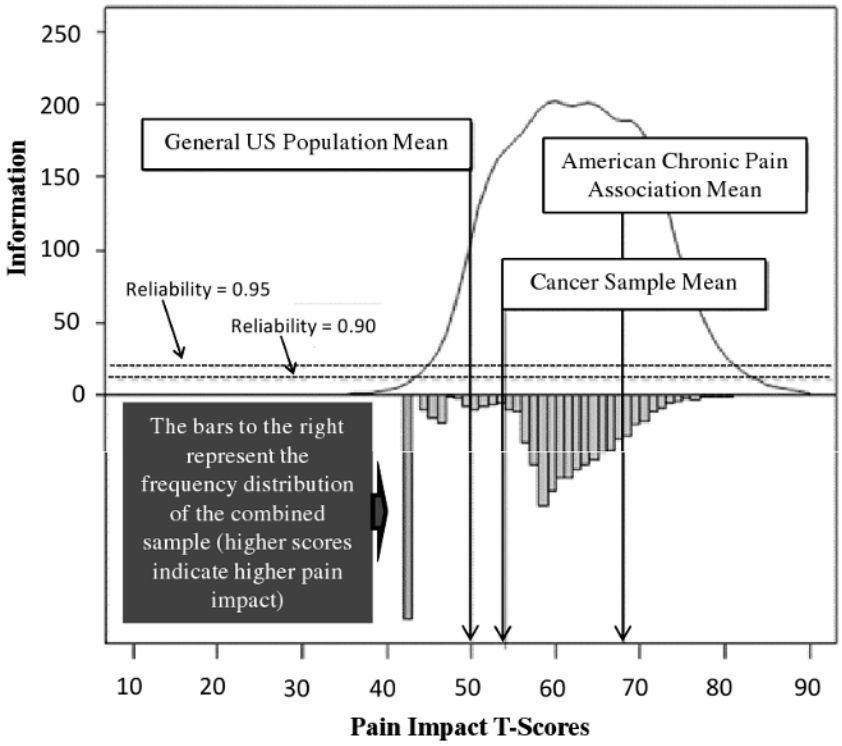

Figure Explanation

- The function plotted on the graph (information) shows how precisely Pain Interference is measured by the PROMIS Pain Interference item bank.
- The amount of information varies for different levels of Pain Interference. As the plot shows, the greatest measurement precision is in the T-Score range of approximately 55 to 70.
- The reference lines for reliabilities of 0.95 and 0.90 show approximately how high the information has to be to have the specified level of reliability (e.g. reliability of 0.90 ≈ information of 10)
- Scores on all PROMIS measures were anchored to a representative US population and have a mean of 50 with a standard deviation of 10.
- The mean of the American Chronic Pain Association (ACPA) sample was 68.2; the mean of the Cancer Sample was 53.1.

Figure 2. Information and Reliability Associated with Pain Impact Items Compared to Distribution of Pain Impact T-Scores in the Combined Samples

Table 1

Comparison of Sample Responses to Emotional and Non-Emotional Pain Interference Items, with Means and Standard Deviations(SD)

	Emotional Items		Non-Emotional Items		Difference Emotional minus Non-Emotional	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Chronic Pain	4.02	0.71	3.79	0.74	0.23	0.6
Cancer	2	0.97	1.85	0.95	0.14	0.47
PROMIS Wave I	1.92	0.97	1.8	0.95	0.1	0.47

Pain. Author manuscript; available in PMC 2011 July 1.

Table 2

Demographics of Calibration Sample by Source

	PROMIS Wave I		Cancer Sample		ACPA	
	N	%	N	%	N	%
RACE/ETHNICITY (participants were allowed to endorse more than one category):						
White	12,293	83	447	84.5	474	91.9
Black/African American	1,199	8.1	78	14.7	10	1.9
Asian	78	0.5	9	1.7	3	0.6
American Indian/Native Alaskan	90	0.6	4	0.8	6	1.2
Native Hawaiian/Pacific Islander	11	0.1	1	0.2	0	0
Other	1,137	7.7	11	2.1	23	4.5
Missing	40	0.3*	3	0.0*	7	1.3*
EDUCATION:						
Less than High School Degree	399	2.7	29	5.3	14	2.7
High School Degree or GED	2,219	15	73	13.9	87	16.7
Some College, Technical School, or Associate Degree	5,531	37.3	159	29.6	239	45.9
College Degree	3,713	25	158	29.9	116	22.3
Advanced Degree	2,976	20.1	112	20.7	65	12.5
Missing	10	0.1*	1	0.2*	2	0.4*
GENDER:						
Male	7,103	47.8	139	26.1	95	18.3
Female	7,744	52.2	390	73.3	425	81.7
Missing	1	0	3	0.6	3	0.6
AGE:						
Mean (SD)	54 (16)		55 (12)		48 (11)	
Range	18-100		18-87		21-86	
Missing	0		0		0	
Total	**14,848**		**532**		**523**	

* Missing percentages based on percent of total cases. All others based on percent of cases reporting.

ACPA—American Chronic Pain Association

Cancer Sample—Collected at NorthShore University HealthSystem and John H. Stroger, Jr. Hospital of Cook County

Table 3

Pearson-Product Moment Correlations (r) and Spearman's Rank Correlations (rho) between PROMIS-Pain Interference Scores and Scores on Other Measures

Measure	N	r
Other PROMIS Measures:		
Physical Function	14,824	-0.55
Fatigue	14,002	0.48
Anxiety	11,911	0.35
Depression	9,558	0.33
Other Pain Measures:		**rho**
BPI Interference Scale	780*	0.9
SF-36 Bodily Pain Subscale	730*	-0.84
0-10 Numerical Rating of Pain Intensity	519**	0.48

* BPI Interference Scale [14] and the SF-36 Bodily Pain Subscale [60] were only administered to participants in the full bank testing arm of the study (N=845).

** The correlation was computed for the chronic pain sample only because of high proportion of people with no pain in the general sample

Table 4

Comparison of PROMIS-Pain Interference (PROMIS-PI) by Number of Chronic and Disabling Conditions, General Health Scores, and Pain Intensity Scores

Variable		ANOVA	Category	N	Mean	SD
Number of Chronic Conditions	F Value (d.f.)	1028 (2)	None	2,071	45.7	6.7
	R-Square	0.12	One	2,682	47.9	7.7
	Pr > F	<0.0001	Two or more	10,095	53.7	9.3
Number of Disabling Conditions	F Value	3179 (2)	None	8,239	47.3	7.2
	R-Square	0.3	One	3,234	54.1	8.2
	Pr > F	<0.0001	Two or more	3,375	59.5	8.6
General Health “In general, would you say your health is excellent/very good/good/fair/poor”	F Value	1646 (4)	Poor	743	63.7	8.7
	R-Square	0.31	Fair	2,704	58.3	8.5
	Pr > F	<0.0001	Good	4,977	52.3	8.3
			Very Good	4,735	47.7	7.2
			Excellent	1,686	44.2	5.4
Pain Intensity “How would you rate your pain on average?” (“0=no pain” through “10=worst imaginable pain”).	F Value	4650 (2)	Mild (0 to 4)	11,113	48.3	7.5
	R-Square	0.39	Moderate (5,6)	2,078	59	6.6
	Pr > F	<0.0001	Severe (7 to 10)	1,650	64.1	6.6