Psychological Test and Assessment Modeling, Volume 58, 2016 (2), 309-352

Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Pain Interference short form items: Application to ethnically diverse cancer and palliative care populations

Jeanne A. Teresi[1,2,3,4], *Katja Ocepek-Welikson*[3], *Karon F. Cook*[5], *Marjorie Kleinman*[2], *Mildred Ramirez*[3,4], *M. Carrington Reid*[4] & *Albert Siu*[6]

Abstract

Reducing the response burden of standardized pain measures is desirable, particularly for individuals who are frail or live with chronic illness, e.g., those suffering from cancer and those in palliative care. The Patient Reported Outcome Measurement Information System® (PROMIS®) project addressed this issue with the provision of computerized adaptive tests (CAT) and short form measures that can be used clinically and in research. Although there has been substantial evaluation of PROMIS item banks, little is known about the performance of PROMIS short forms, particularly in ethnically diverse groups. Reviewed in this article are findings related to the differential item functioning (DIF) and reliability of the PROMIS pain interference short forms across diverse sociodemographic groups.

Methods: DIF hypotheses were generated for the PROMIS short form pain interference items. Initial analyses tested item response theory (IRT) model assumptions of unidimensionality and local independence. Dimensionality was evaluated using factor analytic methods; local dependence (LD) was tested using IRT-based LD indices. Wald tests were used to examine group differences in

[1] *Correspondence concerning this article should be addressed to:* Jeanne A. Teresi, Ed.D, Ph.D., Columbia University Stroud Center at New York State Psychiatric Institute, 1051 Riverside Drive, Box 42, Room 2714, New York, New York, 10032-3702, USA; email: Teresimeas@aol.com

[2] New York State Psychiatric Institute

[3] Research Division, Hebrew Home at Riverdale; RiverSpring Health

[4] Division of Geriatrics and Palliative Medicine, Weill Cornell Medical College

[5] Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University

[6] Brookdale Department of Geriatrics and Palliative Medicine, Mount Sinai School of Medicine

IRT parameters, and to test DIF hypotheses. A second DIF-detection method used in sensitivity analyses was based on ordinal logistic regression with a latent IRT-derived conditioning variable. Magnitude and impact of DIF were investigated, and reliability and item and scale information statistics were estimated.

Results: The reliability of the short form item set was excellent. However, there were a few items with high local dependency, which affected the estimation of the final discrimination parameters. As a result, the item, "How much did pain interfere with enjoyment of social activities?" was excluded in the DIF analyses for all subgroup comparisons.

No items were hypothesized to show DIF for race and ethnicity; however, five items showed DIF after adjustment for multiple comparisons in both primary and sensitivity analyses: ability to concentrate, enjoyment of recreational activities, tasks away from home, participation in social activities, and socializing with others. The magnitude of DIF was small and the impact negligible. Three items were consistently identified with DIF for education: enjoyment of life, ability to concentrate, and enjoyment of recreational activities. No item showed DIF above the magnitude threshold and the impact of DIF on the overall measure was minimal. No item showed gender DIF after correction for multiple comparisons in the primary analyses. Four items showed consistent age DIF: enjoyment of life, ability to concentrate, day to day activities, and enjoyment of recreational activities, none with primary magnitude values above threshold. Conditional on the pain state, Spanish speakers were hypothesized to report less pain interference on one item, enjoyment of life. The DIF findings confirmed the hypothesis; however, the magnitude was small.

Using an arbitrary cutoff point of theta (θ) $\geq$ 1.0 to classify respondents with acute pain interference, the highest number of changes were for the education groups analyses. There were 231 respondents (4 % of the total sample) who changed from the designation of no acute pain interference to acute interference after the DIF adjustment. There was no change in the designations for race/ethnic subgroups, and a small number of changes for respondents aged 65 to 84.

Conclusions: Although significant DIF was observed after correction for multiple comparisons, all DIF was of low magnitude and impact. However, some individual-level impact was observed for low education groups. Reliability estimates were high. Thus, the PROMIS short form pain items examined in this ethnically diverse sample performed relatively well; although one item was problematic and removed from the analyses. It is concluded that the majority of the PROMIS pain interference short form items can be recommended for use among ethnically diverse groups, including those in palliative care and with cancer and chronic illness.

Key words: Differential item functioning, PROMIS®, pain, measurement equivalence, palliative care, ethnicity, cancer

Background

Pain is estimated to affect between 2% and 55% of the population worldwide (Johannes, Le, Zhou, Johnston, & Dworkin, 2010); however, most estimates for chronic pain range from 22 % to 31 % (Breivik, Collett, Ventafridda, Cohen, & Gallacher, 2006; Johannes et al.; Moulin, Clark, Speechley, & Morley-Forster, 2002) and generally increase with age (Johannes et al.). Short standardized assessments of pain are desirable, given the burden of assessment, particularly in frail populations with chronic illness, those suffering from cancer, and those receiving palliative care. Short assessments are highly desirable for clinicians who frequently must assess for and manage multiple conditions in a given visit, particularly among older patients. The aim of the Patient Reported Outcome Measurement Information System® (PROMIS®) project was to provide computerized adaptive tests (CAT) and short form measures that can be used clinically and in research (Reeve et al., 2007). However, little information about measurement equivalence, particularly across ethnically diverse groups is available. The aim of this article is to examine the differential item functioning (DIF) and reliability of the PROMIS short forms across diverse socio-demographic groups using item response theory (IRT) methods.

There are as many as 11 response categories for some pain items, e.g., the Numeric Rating Scale (NRS; Cleeland, 1989). Question arises as to the validity and efficiency of numerous response categories, and some research using advanced psychometric analyses with IRT suggests a more parsimonious categorical representation of pain response (Chen, Revicki, Lai, Cook, & Amtmann, 2009). For example, Orlando-Edelen and Saliba (2010) showed that some of the 11 categories of the NRS items were overlapping and did not provide unique information. Other research (Waterman et al., 2010) confirmed the category overlap in measures with 11 response categories, and provided evidence that the response categories were not interval. Thus fewer categories may be more efficient. For example, Walton, Wideman, and Sullivan (2013) using a Rasch IRT approach found few instances of disordered threshold parameters in pain items with lower numbers (five) of response categories, suggesting a continuous interval level scale. The use of fewer response categories was the approach taken by PROMIS investigators who developed pain items with five response categories (Amtmann, et al., 2010; Cella et al., 2007).

Some efforts to develop and examine pain item banks have included analyses of DIF; e.g., back pain (Kopec et al., 2008) and pain interference in PROMIS (Amtmann et al., 2010). In a study by Amtmann et al. (2010), gender DIF was observed for the pain interference item, enjoyment of life. Age-related DIF was identified for eight items. Among the items examined by Amtmann and colleagues, only the item, ability to concentrate was included in the short form items examined in this paper. In the study by Kopec and colleagues (2008), DIF was observed for two items: pain expression and frustrated in an observational measure of pain suitable for patients with dementia (van Nispen tot Pannerden et al., 2009). Nurses had greater tendencies to report pain expression as present in residents with mild to moderate as contrasted with severe dementia; in contrast they were more likely to report “frustrated” among individuals with severe to very severe dementia as contrasted with milder dementia.

Methods

Sample

These data are from individuals with cancer who were selected from cancer registries in four regions of the United States. Details are provided in the overview article on the sample characteristics (Jensen et al., 2016). The overall sample sizes were 1,053 Hispanics, 917 Asians/Pacific Islanders, 1,122 Blacks and 2,278 non-Hispanic Whites. The focal group was males (n = 2,192) in the analyses of gender; there were 3,247 females. In the analyses of education, the reference group was graduate degree (n = 644). The studied groups were less than high school (n = 965), high school (n = 1,050), some college (n = 1,759), and college degree (n = 983). The reference group for age was 21 to 49 (n = 1,200), the studied groups were 50 to 64 (n = 2,010) and 65 to 84 (n = 2,229). For the analyses of ethnicity the reference group was non-Hispanic Whites (n = 2,269); the studied groups were Blacks (n = 1,119), Hispanics (n = 1,043), and Asians/Pacific Islanders (n = 908). Within the Hispanic sub-sample, there were 334 interviews conducted in Spanish and 704 in English. The majority of the sample had breast, prostate, or colorectal cancer; 17 % were stage III and 12 % stage IV; 40 % reported two or more comorbidities (Jensen et al., 2016).

Measure

The PROMIS pain interference short form items were selected based on provision of maximum information as well as clinical review to represent both the content of the bank and precision across the range of the trait measured. Ten items from four short forms were examined: 4a, 6a, 6b, and 8a (see Table 1 for content). An eleventh global pain rating item was included in the qualitative analyses, but was not examined as part of the quantitative analyses.

Analyses

Qualitative analyses and hypotheses generation

Generation of DIF Hypotheses: Qualitative analyses were conducted originally (DeWalt, Rothrock, Yount, & Stone, 2007) to ensure that the meaning of items was clear. DIF hypotheses were generated for this study by asking a set of clinicians and other content experts to indicate whether or not they expected DIF to be present, and the direction of the DIF with respect to several comparison groups: gender, age, race/ethnicity, language, education, and diagnosis. They provided the hypotheses in terms of presence and direction of DIF. The goal was to identify items that might have a different meaning or not be understood well and/or equivalently by individuals of any of the groups referenced. A grid containing a row for each of the eleven items and separate columns for each of the referenced groups was distributed to the experts for completion in order to facilitate the rating (see Table 1).

Table 1:
Summary of pain interference DIF hypotheses generated by eight content experts

STEM: How much did pain interfere …	GENDER	AGE	RACE/ ETHNIC	LANGUAGE	EDUCATION	DX
With your enjoyment of life? *(Short forms 8a, 6b)*	4[a] Women more interference (3)[b]	2 Older more interference		2 Non-English speakers less interference; Spanish less interference		
With your ability to concentrate? *(Short form 6b)*	4 Women more interference(2)	5 Older more interference(3)			2 Lower education more interference	2
With your day to day activities? *(Short forms 4a, 6a, 6b, 8a)*	4	6 Older more interference(4)			2 Lower education more interference	3
With your enjoyment of recreational activities?*(Short form 6b)*	2	3 Older more interference				2
With doing your tasks away from home (e.g., getting groceries, running errands)?*(Short form 6b)*	4	4 Younger more interference (2)				2
With your ability to participate in social activities? *(Short form 4a, 6a, 8a)*	4 Women more interference(2)	2 Older more interference				
With work around the home? (*Short form 4a, 6a, 8a)*	6 Women more interference (3)	5 Younger more interference(3)				2
With the things you usually do for fun?*(Short form 6a, 8a)*						2
With your enjoyment of social activities?*(Short form 6a, 8a)*	4 Women more interference(3)					
How often did pain keep you from socializing with others? *(Short form 6b)*	4 Women more interference(2)					
How would you rate your pain on average?	2 Males higher levels of pain					

Note: Direction is given for those with two or more consistent ratings; short form items are provided in parentheses after the item wording.
a Number indicates total number of hypotheses; b Number indicates number of directional hypotheses if not the same as the number rating (a).
Dx=diagnosis

Definition of DIF: A definition of DIF was provided, and the following instructions given.

Differential item functioning means that individuals in groups with the same underlying trait (state) level will have different probabilities of endorsing an item. Put another way, reporting interference (e.g., in day to day activities) associated with pain should depend only on the level of the trait (state), e.g., pain, and not on membership in a group, male or female. Very specifically, randomly selected persons from each of two groups (e.g., males and females) who are at the same (e.g., mild) level of pain interference should have the same likelihood of reporting interference in day to day activities, resulting from pain. If it is theorized that this might not be the case, it would be hypothesized that the item has gender DIF.

Quantitative analyses

Model Assumption of Unidimensionality: Unidimensionality was examined using split samples, constructed by selection of two random halves in order to use one sample for cross-validation of results. The first random half of a split sample was used to perform exploratory principal components analyses and to fit a unidimensional confirmatory factor analysis (CFA). Essential unidimensionality was examined through a merged exploratory factor analysis (EFA) and CFA (Asparouhov & Muthén, 2009) performed by fitting a unidimensional model with polychoric correlations using Mplus (Muthén & Muthén, 2011).

The confirmatory analyses of the unidimensional model and evaluation of the Comparative Fit Index (CFI) was performed in the context of invariance testing and model fit (Bentler, 1990; Cook, Kallen, & Amtmann, 2009; Meade, Johnson, & Bradley, 2008). A bifactor model was compared with a unidimensional model. A Schmid-Leiman (S-Lplus; Schmid & Leiman, 1957) transformation using the "psych" R package (Rizopoulus, 2009) was performed with the second random half of the sample in order to find an alternative set of group factors for the bi-factor model (Reise, Moore, & Haviland, 2010). All items were specified to load on the general factor, and the loadings on the group factors were specified following the Schmid-Leiman solution. Mplus (Muthén & Muthén, 2011) was used to both estimate the polychoric correlations based on the underlying continuous normal variables and to perform the final bi-factor modeling.

The explained common variance (ECV) provides information about whether the observed variance covariance matrix is close to unidimensionality (Sijtsma, 2009), and is estimated as the percent of observed variance explained (Reise, 2012; Reise et al., 2010).

Local Dependence (LD): The generalized, standardized local dependency χ^2statistics (Chen & Thissen, 1997) provided in IRTPRO, version 2.1 (Cai, Thissen, & du Toit, 2011) were used to test the local independence assumption. Sensitivity analyses removing one item each from two pairs of items with higher LD values was performed.

IRT-model Fit: Model fit for the IRT model was examined using the root mean square error of approximation (RMSEA) from IRTPRO (Cai et al., 2011).

Anchor Items and Linking: An iterative process was used in selection of the DIF-free anchor items for θ estimation. The method that was used in these analyses is a modified "all-other" anchor method in which initial DIF estimates can be obtained by treating each item as a "studied" item, while using the remainder as "anchor" items (see Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006). The analyses were repeated using the final subset of items identified as free of DIF as the "purified" anchor set. Items with DIF from the original anchor set were removed.

Model for DIF Detection: The graded response model (Samejima, 1969) was used for the analyses of DIF. A nominal response model was also examined in sensitivity analyses. The formula is given in the methods overview to this series (Teresi & Jones, 2016). The item characteristic curve (ICC) that relates the probability of an item response to the underlying state, e.g., pain interference, measured by the item set is characterized by: a discrimination parameter, proportional to the slope of the curve (denoted *a*) and location (severity) parameter(s) (denoted *b*). An item shows DIF if people from different subgroups but at the same level of the attribute (denoted θ) have unequal probabilities of endorsement. The presence of DIF is demonstrated by ICCs that are different for the subgroups examined.

DIF Detection Tests: The primary method used for DIF detection was the Wald test for examination of group differences in IRT item parameters. For each studied item, a model was constructed with all parameters (except the studied item) constrained to be equal across comparison groups for the anchor items, and item parameters for the studied item freed to be estimated distinctly. An overall simultaneous joint test of differences in the *a* or *b* parameters was performed followed by step down tests for group differences in the *a* parameters, followed by conditional tests of the *b* parameters. Uniform DIF is detected when the *b* parameters differ and non-uniform DIF when the *a* parameters differ.

Because there were three or more groups (three age, four race/ethnicity, and five education), and the interest is in comparing the studied groups to the reference group, non-orthogonal rather than orthogonal contrasts were used. The final *p* values were adjusted using Bonferroni (Bonferroni, 1936) methods. IRTPRO option 3 (Cai et al., 2011) was used for DIF detection.

Sensitivity Analyses for DIF Detection: A second DIF-detection method used in sensitivity analysis was based on ordinal logistic regression (OLR; Swaminathan & Rogers, 1990; Zumbo, 1999), which typically conditions on an observed variable. Uniform DIF is defined in the OLR framework as a significant group effect, conditional on the pain state; non-uniform DIF is a significant interaction of group and state. Three hierarchical models were tested; the first examines pain state (1), followed by group (2), and the interaction of group by state (3). Non-uniform DIF was tested by examining model 3 vs. 2; then uniform DIF was tested by examining the incremental effect of model 2 vs. 1, with a χ^2(1 degree of freedom) test (Camilli, & Shepard, 1994). A modification applied in these analyses, IRTOLR (Crane, Gibbons, Jolley, & van Belle, 2006; Crane, van Belle, & Larson, 2004; Mukherjee, Gibbons, Kristiansson, & Crane, 2013) uses the pain interference estimates from a latent variable IRT model rather than the traditional observed score conditioning variable, and incorporates effect sizes into the uniform DIF

detection procedure. The software lordif (Choi, Gibbons, & Crane, 2011) was used to perform IRTOLR. Details of the procedure are given in the methods overviews in this series (Kleinman & Teresi, 2016; Teresi & Jones, 2016).

Evaluation of DIF Magnitude and Effect Sizes: Expected item scores were examined as measures of magnitude. (See Figure 1 for examples of graphs of expected scale and item score functions for comparison groups.) An expected item score is the sum of the weighted (by the response category value) probabilities of scoring in each of the possible categories for the item. A method for quantification of the difference in the average expected item scores is the non-compensatory DIF index (NCDIF; Raju, van der Linden, & Fleer, 1995) used in Differential Functioning of Items and Tests (DFIT; Flowers, Oshima, & Raju, 1999; Oshima, Kushubar, Scott, & Raju, 2009; Raju, 1999; Raju et al., 2009). For the data presented here, the cutoff values are 0.0960 for polytomous items with five response options (Raju, 1999). Additional effect size measures proposed by Wainer (1993) and extended for polytomous data by Kim, Cohen, Alagoz, and Kim (2007) were also examined. For example, also reported here is the *T1* effect size measure (Wainer, 1993), for which a recommended cutoff value is 0.10. However, primary reliance was on the NCDIF magnitude measure because little research has been conducted on the performance of *T1*. For a detailed description of these measures see Kleinman and Teresi (2016). For the sensitivity analyses using IRTOLR, the pseudo-R^2 measures of Cox and Snell (1989), Nagelkerke (1991), and McFadden (1974) were used to examine magnitude of DIF.

Evaluation of DIF Impact: Aggregate-level impact was evaluated, examining expected scale score functions. Expected item scores were summed to produce an expected scale score (also referred to as the test or scale response function), which provides evidence regarding the effect of DIF on the total score. Group differences in these test (scale) response functions provide overall aggregated measures of DIF impact.

Impact at the individual level was examined by comparing DIF-adjusted and unadjusted estimates of the latent pain state scores. Individual impact was evaluated by fixing and freeing parameters to account for DIF, and comparing the results with and without DIF adjustment. Individual impact was presented in two ways: 1) the number of individual θ estimates that differ by more than 0.5 or 1.0 standard deviations; 2) based on a threshold value.

Evaluation of Reliability and Information: McDonald's (McDonald, 1999) omega total (ω_t), a reliability estimate that is based on the proportion of total common variance explained was calculated. Additionally, IRT-based reliability measures were examined at selected points along the underlying latent continuum. Finally, the item and scale information functions from IRT were calculated and graphed.

Results

Qualitative results

The pain items were reviewed qualitatively by eight content experts regarding potential sources of differential item functioning. Three of the members of the panel were clinical

or counseling psychologists, three were public health professionals, and two were gerontologists.

Gender DIF was posited for 10 out of the 11 pain interference items, seven with directionality, of which six suggested that women as contrasted with men will be more likely to report greater interference (due to pain) with enjoyment of life and of social activities, ability to concentrate, ability to participate in social activities, work around the home, and with socializing with others. Similarly, age DIF was posited with directionality for seven out of the eleven items, suggesting that conditional on pain level, older individuals would be more likely than younger individuals to endorse responses indicative of higher levels of interference due to pain for the items: enjoyment of life and recreational activities, ability to concentrate and day to day and social activities. Younger individuals (in contrast to older individuals) were posited to express more pain interference in the areas of doing tasks at home and away from home. There were no race/ethnicity DIF hypotheses posited. Language DIF was posited for an item suggesting that non-English speakers and Spanish speakers would be less likely to report pain-related interference with enjoyment of life. Similarly, with respect to education DIF, it was suggested that conditional on pain, those individuals with lower levels of education will be more likely to endorse responses indicating greater interference with ability to concentrate, and with day-to-day activities (see Table 1).

Quantitative results

Distributions

As expected, skew was observed for all items in the direction of a smaller proportion (4 % to 10 %) responding in the most extreme category, *very much*. The majority (43 % to 57 %) reported *not at all*. Because of the large sample sizes, there were few instances of sparse data.

Tests of model assumptions

Unidimensionality: As shown in Table 2, there was strong support for essential unidimensionality across all comparison socio-demographic groups. (The test of scree for the total sample is given in Appendix Figure A1.) The principal components analyses showed that the ratio of component 1 to 2 was very large (32.1 to 49.1) across groups. The first component accounted for between 90 % and 93 % of the variance across groups, supporting the essential unidimensionality of the item set across comparison subgroups (see Table 2). As an additional test of dimensionality a bifactor model was examined using the second random half of the sample.

Table 2:
PROMIS pain interference ten item set: Tests of dimensionality from principal components analysis (eigenvalues by subgroup)

Statistic	Component 1	Component 2	Component 3	Component 4	Ratio Component 1/ Component 2
Total Sample (n = 5,475)					
Eigenvalues	9.189	0.222	0.156	0.110	41.4
Explained Variance	91.9%	2.2%	1.6%	1.1%	
Random First Half Sample (n = 2,737)					
Eigenvalues	9.216	0.211	0.148	0.104	43.7
Explained Variance	92.2%	2.1%	1.5%	1.0%	
Females (n = 3,247)					
Eigenvalues	9.167	0.239	0.161	0.110	38.4
Explained Variance	91.7%	2.4%	1.6%	1.1%	
Males (n = 2,192)					
Eigenvalues	9.223	0.199	0.149	0.107	46.3
Explained Variance	92.2%	2.0%	1.5%	1.1%	
Age 21-49 (n = 1,200)					
Eigenvalues	9.279	0.189	0.153	0.093	49.1
Explained Variance	92.8%	1.9%	1.5%	0.9%	
Age 50-64 (n = 2,010)					
Eigenvalues	9.238	0.212	0.137	0.097	43.6
Explained Variance	92.4%	2.1%	1.4%	1.0%	
Age 65-84 (n = 2,229)					
Eigenvalues	9.072	0.273	0.177	0.132	33.2
Explained Variance	90.7%	2.7%	1.8%	1.3%	
Race/Ethnicity: Non-Hispanic White (n = 2,269)					
Eigenvalues	9.237	0.214	0.159	0.083	43.2
Explained Variance	92.4%	2.1%	1.6%	0.8%	
Race/Ethnicity: Black (n = 1,119)					
Eigenvalues	9.142	0.241	0.147	0.128	37.9
Explained Variance	91.4%	2.4%	1.5%	1.3%	
Race/Ethnicity: Hispanic (n = 1,043)					
Eigenvalues	9.116	0.258	0.165	0.112	35.3
Explained Variance	91.2%	2.6%	1.6%	1.1%	

Statistic	Component 1	Component 2	Component 3	Component 4	Ratio Component 1/ Component 2
Race/Ethnicity: Asian/ Pacific Islander (*n* = 908)					
Eigenvalues	9.192	0.240	0.166	0.107	38.3
Explained Variance	91.9%	2.4%	1.7%	1.1%	
Education: Less Than High School (*n* = 965)					
Eigenvalues	9.011	0.277	0.188	0.143	32.5
Explained Variance	90.1%	2.8%	1.9%	1.4%	
Education: High School (*n* = 1,050)					
Eigenvalues	9.161	0.263	0.158	0.116	34.8
Explained Variance	91.6%	2.6%	1.6%	1.2%	
Education: Some College (*n* = 1,759)					
Eigenvalues	9.217	0.210	0.150	0.104	43.9
Explained Variance	92.2%	2.1%	1.5%	1.0%	
Education: College Degree (*n* = 983)					
Eigenvalues	9.140	0.277	0.170	0.118	33.0
Explained Variance	91.4%	2.8%	1.7%	1.2%	
Education: Graduate Degree (*n* = 644)					
Eigenvalues	9.265	0.202	0.155	0.095	45.9
Explained Variance	92.6%	2.0%	1.5%	0.9%	
Hispanics Interviewed in English (*n* = 704)					
Eigenvalues	9.146	0.252	0.155	0.108	36.3
Explained Variance	91.5%	2.5%	1.6%	1.1%	
Hispanics Interviewed in Spanish (*n* = 334)					
Eigenvalues	9.078	0.283	0.181	0.109	32.1
Explained Variance	90.8%	2.8%	1.8%	1.1%	

Examination of the confirmatory factor analyses results in Table 3 shows that the loadings on the single common factor were very similar to those observed on the general factor from the bifactor analyses (the differences in the final Mplus loadings and those from the one factor solution were between 0 to 0.03), which provides additional evidence for unidimensionality. Additionally, the communality values were large, ranging from 0.83 to 0.97 for the two group model. There were no loadings ≥ 0.20 on the 3rd group factor in the 3 group model. The model fit indices (CFIs) for the unidimensional CFA from Mplus ranged from 0.997 to 0.999 across groups (see Appendix Table A1); the ECVs ranged from 84.94 to 88.14 (see Table 4).

Table 3:

PROMIS pain interference ten item set: Item loadings (λ) from the unidimensional confirmatory factor analysis (Mplus) for the first half of the random sample (n = 2,737), Schmid-Leiman bi-factor model with two and three group factors performed with R and Mplus bi-factor two group solution for the second random half of the sample (n = 2,738)

Item Description	One Fact.* λ (s.e.)	Schmid-Leiman Bi-Factor Three and Two Group Factor Solutions									MPLUS Bi-Factor Solution (Based on S-L** Result) #		
		G λ	F1 λ	F2 λ	h^2	G λ	F1 λ	F2 λ	F3 λ #	h^2	G λ (s.e.)	F1 λ (s.e.)	F2 λ (s.e.)
How much did pain interfere with your enjoyment of life?	0.95 (0.002)	0.94		0.20	0.92	0.93		0.25		0.92	0.93 (0.003)		0.25 (0.015)
How much did pain interfere with your ability to concentrate?	0.92 (0.004)	0.90			0.83	0.89		0.21		0.84	0.89 (0.005)		0.22 (0.016)
How much did pain interfere with your day to day activities?	0.97 (0.002)	0.96			0.95	0.95		0.23		0.96	0.96 (0.002)		0.15 (0.010)
How much did pain interfere with your enjoyment of recreational activities?	0.96 (0.002)	0.95			0.93	0.94				0.94	0.97 (0.002)		
How much did pain interfere with doing your tasks away from home?	0.97 (0.002)	0.95			0.94	0.94	0.20			0.93	0.97 (0.002)	0.09 (0.009)	
How much did pain interfere with your ability to participate in social activities?	0.97 (0.001)	0.96	0.20		0.96	0.95	0.24			0.97	0.96 (0.002)	0.17 (0.010)	
How much did pain interfere with work around the home?	0.97 (0.002)	0.95			0.93	0.94				0.93	0.98 (0.002)		
How much did pain interfere with the things you usually do for fun?	0.97 (0.002)	0.96			0.96	0.95	0.21			0.96	0.96 (0.002)	0.15 (0.010)	
How much did pain interfere with your enjoyment of social activities?	0.97 (0.002)	0.96	0.21		0.97	0.95	0.24			0.97	0.96 (0.002)	0.22 (0.011)	
How often did pain keep you from socializing with others?	0.93 (0.004)	0.91			0.85	0.90	0.20			0.86	0.91 (0.005)	0.20 (0.014)	

* Geomin (oblique) rotation ** Schmid-Leiman bi-factor model # No items loaded on the 3rd group factor ≥ 0.20; the Mplus solution excluded the third non-loading factor from the three group factor solution;

$h^2$is the communality. G λ are the loadings on the general factor; F1 λ through F3 λ are the loadings on the group factors

Note: Comparative fit index (CFI) for the Mplus one-factor solution is 0.998 and for the bi-factor solution is 0.999

Table 4:
PROMIS short form pain ten (and nine) item sets. Reliability statistics: Cronbach's and ordinal alpha, McDonald's Omega Total and explained common variance (ECV) for the total sample and demographic subgroups ("Psych" R package)

	Cronbach's Alpha	Ordinal Alpha	McDonald's Omega	ECV
Total Sample	0.983 (0.980)	0.990 (0.989)	0.991 (0.989)	86.904 (86.528)
Random Second Half Sample	0.983 (0.980)	0.990 (0.988)	0.990 (0.988)	86.581 (86.185)
Age 21 to 49 years	0.985 (0.983)	0.991 (0.990)	0.992 (0.990)	88.144 (87.902)
Age 50 to 64 years	0.985 (0.982)	0.991 (0.990)	0.991 (0.990)	87.854 (87.555)
Age 65 to 84 years	0.980 (0.977)	0.989 (0.987)	0.989 (0.987)	84.937 (84.387)
Male	0.983 (0.981)	0.991 (0.989)	0.991 (0.989)	87.123 (86.686)
Female	0.983 (0.980)	0.990 (0.988)	0.990 (0.989)	86.744 (86.409)
Non-Hispanic White	0.983 (0.980)	0.991 (0.989)	0.991 (0.990)	86.850 (86.390)
Non-Hispanic Black	0.983 (0.980)	0.990 (0.988)	0.990 (0.988)	86.551 (86.274)
Hispanic	0.983 (0.980)	0.989 (0.987)	0.989 (0.988)	86.423 (86.106)
Non-Hispanic Asian/ Pacific Islander	0.983 (0.980)	0.990 (0.989)	0.991 (0.989)	86.782 (86.404)
Less Than High School	0.981 (0.978)	0.988 (0.986)	0.988 (0.986)	85.386 (85.039)
High School Degree	0.983 (0.980)	0.990 (0.988)	0.990 (0.988)	86.611 (86.228)
Some College	0.984 (0.981)	0.991 (0.989)	0.991 (0.989)	87.314 (86.956)
College Graduate	0.980 (0.977)	0.989 (0.988)	0.990 (0.988)	84.996 (84.482)
Graduate Degree	0.982 (0.979)	0.991 (0.990)	0.991 (0.990)	86.059 (85.679)
Hispanics Interviewed in English	0.983 (0.980)	0.990 (0.989)	0.990 (0.989)	86.708 (86.595)
Hispanics Interviewed in Spanish	0.982 (0.979)	0.989 (0.987)	0.989 (0.987)	85.941 (85.707)

Local Independence: In general, the local dependency statistics (not shown) were in the acceptable range. However, there were a few high LD statistics, the highest between item 9 – "How much did pain interfere with enjoyment of social activities?" and item 8 – "How much did pain interfere with the things you usually do for fun?" (LD = 28.3 for the low education subgroup and 25.6 for the Black subgroup) when the 10 item set was analyzed. The local dependency among items affected the estimation of the final *a* parameters; for example in the race/ethnicity DIF analysis, the estimates for items 8 and 9 were above 10.0 for all subgroups (not shown) and were above 10.0 in the individual group IRT analyses for more than one third of the analyses (see Appendix Table A2). As a result, item 9 was excluded in the IRTDIF analyses for all subgroup comparisons, which improved the parameter estimation.

Table 5:
PROMIS pain interference nine item set: Item response theory (IRT) reliability estimates at varying levels of the attribute *(θ)* based on the results of the IRT analysis (IRTPRO) for the total sample and demographic subgroups

Pain Interference (θ)	IRT Reliability																
	Total	**F**	**M**	**Age 21-49**	**Age 50-64**	**Age 65-84**	**NHW**	**NHB**	**Hisp.**	**NH API**	**<HS**	**HS**	**Some Coll.**	**Coll.**	**Grad.**	**Lang. Engl.**	**Lang. Span.**
-1.2	0.56	0.64	0.55	0.60	0.63	0.56	0.53	0.66	0.68	0.57	0.84	0.65	0.62	0.52	0.51	0.68	0.70
-0.8	0.78	0.90	0.74	0.88	0.90	0.76	0.70	0.91	0.92	0.81	0.97	0.90	0.89	0.64	0.56	0.92	0.92
-0.4	0.96	0.99	0.95	0.99	0.99	0.95	0.94	0.98	0.99	0.97	0.99	0.99	0.98	0.89	0.83	0.99	0.99
0.0	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.99
0.4	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
0.8	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
1.2	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99
1.6	0.99	0.98	0.99	0.97	0.97	0.99	0.99	0.97	0.97	0.99	0.94	0.97	0.98	0.99	0.99	0.97	0.97
2.0	0.95	0.90	0.95	0.89	0.89	0.96	0.95	0.90	0.89	0.95	0.82	0.90	0.91	0.98	0.98	0.89	0.90
2.4	0.83	0.74	0.83	0.72	0.70	0.87	0.86	0.75	0.71	0.79	0.65	0.74	0.74	0.93	0.92	0.70	0.75
Overall (Average)	**0.90**	**0.91**	**0.90**	**0.90**	**0.90**	**0.91**	**0.89**	**0.91**	**0.91**	**0.90**	**0.92**	**0.91**	**0.91**	**0.89**	**0.87**	**0.91**	**0.92**

Note: Reliability estimates are calculated for θ levels for which there are respondents

Tests of model fit

The fit statistics (RMSEAs) from IRTPRO for the IRT models (see Appendix Table A1) ranged from 0.05 to 0.11 for the ten item set and from 0.05 to 0.10 for the nine item set across DIF subgroup comparison models, indicating good to acceptable fit. Across analyses, the fit for the nine item set improved by 0.01 to 0.02. The RMSEA for the nine item set ranged from 0.05 to 0.07 for all analyses, except for the subgroup of Hispanics interviewed in English. The fit of the nominal response model tested in sensitivity analyses was poor.

Reliability estimates

The reliability estimates were high. The omega total values for the nine item scale (Table 4) ranged from 0.986 to 0.990, the Cronbach's alphas ranged from 0.977 to 0.983, and ordinal alphas based on polychoric correlations from 0.986 to 0.990. Finally, the reliability estimates (precision) at points along the latent trait (θ) reflective of where respondents were observed were high. As shown in Table 5, the overall reliability estimate was 0.90 for the total sample ranging from 0.87 to 0.92 for the individual subgroups. However there was some variability among the subgroups in terms of reliability estimates at different θ levels. The reliability was low at the lowest point (θ = -1.2), ranging from 0.51 (respondents with graduate education) to 0.84 (less than high school). Low reliability extended to the next θ level (-0.8) for the respondents with graduate education (0.56) and those with a college degree (0.64). The reliability was also lower (0.65) for the respondents with less than high school education at θ level 2.4. The overall reliability estimates for the range of θ from -1.2 to 2.4 (where most respondents scored) were from 0.87 to 0.92 (see Table 5).

IRT parameter estimates, tests of DIF and assessment of magnitude and impact

Shown in Table 6 are the graded response item parameters and their standard errors for the total sample for the nine item set. Appendix Tables A2 and A3 show the discrimination (a) parameters across subgroup comparisons for the ten and nine item sets, respectively. Comparatively, the estimates of the a parameters in the ten item set analyses were higher overall, ranging from 3.44 (pain keeps you from socializing with others) for respondents interviewed in Spanish to 12.19 (pain interference in enjoyment of social activities) for respondents with a graduate education. Estimates of the discrimination parameters were from 7.90 to 12.19 among individual subgroups for item 8, interference with things you do for fun, and item 9, pain interference in enjoyment of social activities (see Appendix Table A2).

As shown, the a parameters although high overall were reduced after removal of the item, enjoyment of social activities, ranging from 4.16 (ability to concentrate) to 8.21 (pain interferes with social activities) across items for the total sample. (See Appendix Table A3.) The pattern was similar for the subgroups with some variation; estimates ranged from 3.34 (pain keeps you from socializing with others) for those interviewed in Spanish to 9.18 (things you do for fun) for those with a high school education.

Table 6:
PROMIS pain interference nine item set: Item response theory (IRT) item parameters and standard error estimates (using IRTPRO) for the total sample (n = 5,475)

Item description	*a*	s.e. of *a*	*b1*	s.e.	*b2*	s.e.	*b3*	s.e.	*b4*	s.e.
How much did pain interfere with your enjoyment of life?	5.25	0.13	-0.14	0.02	0.51	0.01	0.98	0.01	1.57	0.02
How much did pain interfere with your ability to concentrate?	4.16	0.10	0.27	0.02	0.79	0.01	1.29	0.02	1.86	0.03
How much did pain interfere with your day to day activities?	6.76	0.18	-0.08	0.02	0.54	0.01	0.95	0.01	1.53	0.02
How much did pain interfere with your enjoyment of recreational activities?	6.47	0.17	-0.11	0.02	0.46	0.01	0.84	0.01	1.36	0.02
How much did pain interfere with doing your tasks away from home?	7.28	0.20	0.09	0.02	0.58	0.01	1.00	0.01	1.49	0.02
How much did pain interfere with your ability to participate in social activities?	8.21	0.25	0.14	0.02	0.60	0.01	1.01	0.01	1.49	0.02
How much did pain interfere with work around the home?	7.73	0.21	-0.07	0.02	0.49	0.01	0.91	0.01	1.47	0.02
How much did pain interfere with the things you usually do for fun?	8.03	0.23	-0.02	0.02	0.50	0.01	0.90	0.01	1.39	0.02
How often did pain keep you from socializing with others?	4.34	0.11	0.13	0.02	0.56	0.01	1.19	0.02	1.90	0.03

DIF tests

Appendix Tables A4 to A8 show the detailed DIF results for race/ethnicity, education, age, gender, and language of interview, respectively while Tables 7 to 10 are summaries of the DIF results after exclusion of one item.

Race and Ethnicity: Table 7 shows summary results for race/ethnicity. No items were hypothesized to show DIF (see Table 1); however, five items showed DIF after Bonferroni correction for analyse performed by both IRTPRO (Wald tests) and by lordif (ordi-

Table 7:
PROMIS pain interference item set: Differential item function (DIF) results. Race/Ethnicity subgroup comparisons

Item description	IRTPRO			lordif			Magnitude (NCDIF)			Effect Size T1		
	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI
How much did pain interfere with your enjoyment of life?	Anchor item			U*		NU; U*	0.0027	0.0017	0.0021	-0.0094	0.0212	0.0270
How much did pain interfere with your ability to concentrate?		U*	NU*; U*	U*	U*	NU*; U*	0.0016	0.0189	0.0544	-0.0066	-0.1042†	-0.1684†
How much did pain interfere with your day to day activities?	NU			U*		NU; U*	0.0015	0.0013	0.0022	0.0286	0.0133	0.0324
How much did pain interfere with your enjoyment of recreational activities?	U	U	U*	NU; U	U*	U*	0.0141	0.0194	0.0345	0.0960	0.1138†	0.1277†
How much did pain interfere with doing your tasks away from home?		U	U*	U*		U*	0.0020	0.0039	0.0035	-0.0283	-0.0271	-0.0423
How much did pain interfere with your ability to participate in social activities?	U*	U*	U*	NU*; U*;	NU; U	U*	0.0083	0.0043	0.0035	-0.0531	-0.0325	-0.0344
How much did pain interfere with work around the home?	Anchor item			NU; U*		U*	0.0008	0.0003	0.0057	0.0168	0.0030	0.0499
How much did pain interfere with the things you usually do for fun?				NU; U*	U*	U*	0.0005	0.0038	0.0103	-0.0096	0.0479	0.0702
How much did pain interfere with your enjoyment of social activities?	Item excluded											
How often did pain keep you from socializing with others?	U*	NU*; U*	U*	NU*; U*	NU; U*	NU; U*	0.0124	0.0176	0.0179	-0.0754	-0.966	-0.0922

All non-compensatory DIF (NCDIF) values were smaller than the threshold (0.0960) *Asterisks indicate significance after adjustment for multiple comparisons. † Indicates value above threshold of 0.10.

NU= Non-uniform DIF involving the discrimination parameters; U=Uniform DIF involving the location parameters.

For the lordif analyses, the uniform and non-uniform DIF was determined using the likelihood ratio χ^2 test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R^2 measures of Cox & Snell (1989), Nagelkerke (1991), and McFadden (1974) or when using the change in Beta criterion.

nal logistic regression). These items were: ability to concentrate, enjoyment of recreational activities, tasks away from home, participation in social activities, and socializing with others. Conditional on the level of pain interference, Asians/Pacific Islanders (as contrasted with non-Hispanic Whites) evidenced a higher probability of responding in the direction of more interference (lower *b* parameters) for five items showing DIF except for one: enjoyment with recreational activities (see Appendix Table A4). The latter item was less likely to be endorsed in the direction of more interference by the Asians/Pacific Islanders as compared to non-Hispanic White respondents. The same result as for the Asians/Pacific Islanders was observed for Hispanics; however, the DIF for the item, tasks away from home was not significant after the correction for multiple comparisons. Conditional on pain interference, only two items showed significant DIF after corrections for multiple comparisons for Black responders (at lower levels of θ) in the direction of more pain interference in participation in social activities and socializing with others than for non-Hispanic Whites (see Appendix Table A4 and Figure 1).

Two items showed DIF of higher magnitude (above threshold on the *T1* statistic) for Hispanics and Asians/Pacific Islanders vs. Whites: ability to concentrate and enjoyment of recreational activities (see Table 7). The magnitude of DIF was small and none of the NCDIF statistics were above threshold. The impact of DIF was negligible, as shown by the overlapping curves (see Figure 1).

Education: The items, ability to concentrate and pain interference with day to day activities were hypothesized to evidence DIF in the direction of more pain interference reported by those with lower education. Three items were identified consistently with DIF for education after Bonferroni correction using the Wald test and OLR: enjoyment of life, ability to concentrate and enjoyment of recreational activities (see Table 8). The item, ability to concentrate, showed non-uniform DIF, and was more discriminating (higher *a* parameter) for respondents with a graduate degree as contrasted with all other educational groups (see Appendix Tables A3 and A5). The item, enjoyment of life, also evidenced non-uniform DIF, and was less discriminating for respondents with less than high school education as contrasted with those with a graduate degree. The item, enjoyment of recreational activities evidenced uniform DIF; those with a graduate degree were less likely to endorse the item in the pain interference direction compared with respondents with less than high school (see Appendix Table A5). No item showed DIF of higher *T1* magnitude and no NCDIF statistics were above threshold (see Table 8).

Figure 1:
PROMIS pain interference nine item set: Expected scale and item score functions for race/ethnicity subgroups

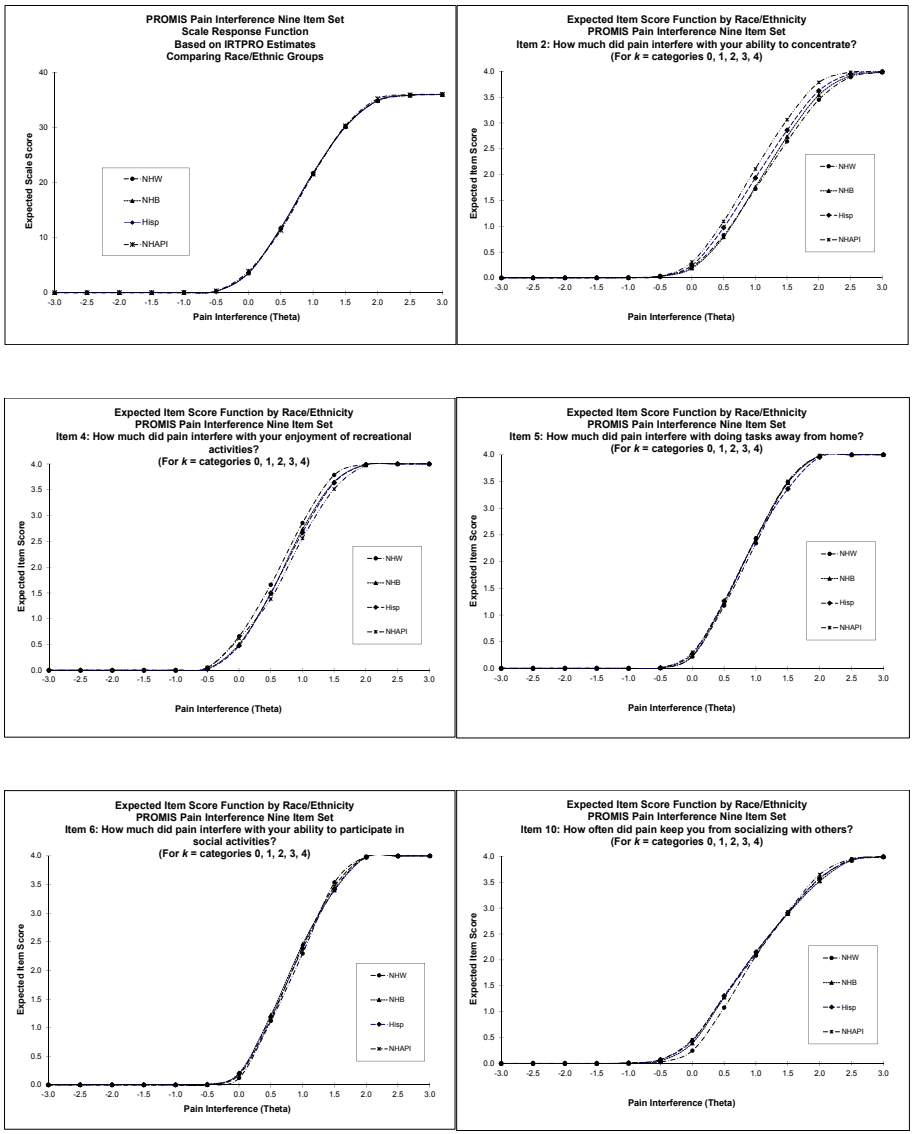
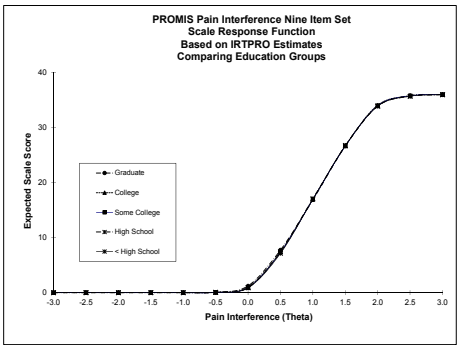
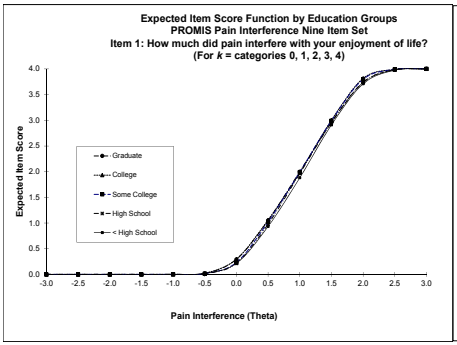
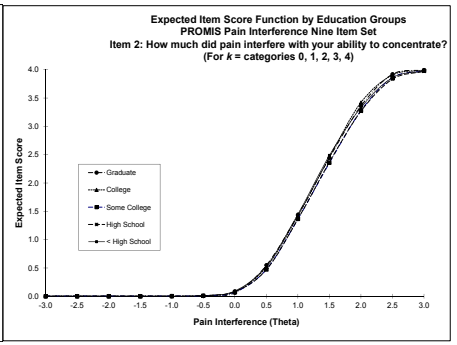


Figure 1 - cont.:
PROMIS pain interference nine item set: Expected scale and item score functions for education subgroups







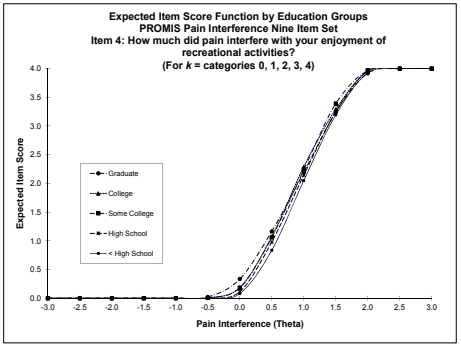
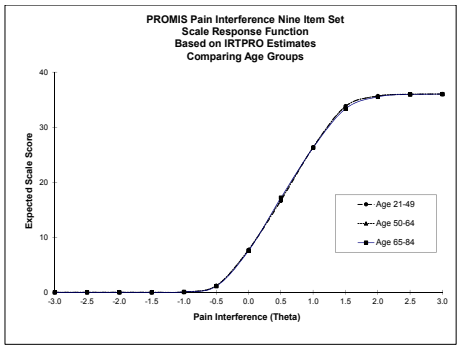
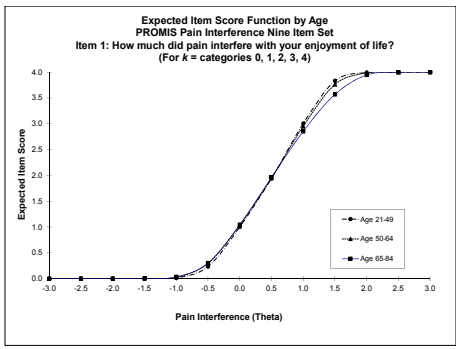
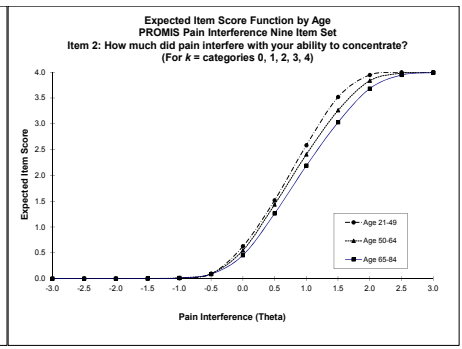
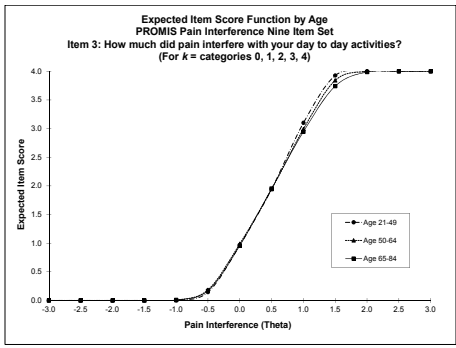


Figure 1 - cont.:
PROMIS pain interference nine item set: Expected scale and item score functions for age subgroups









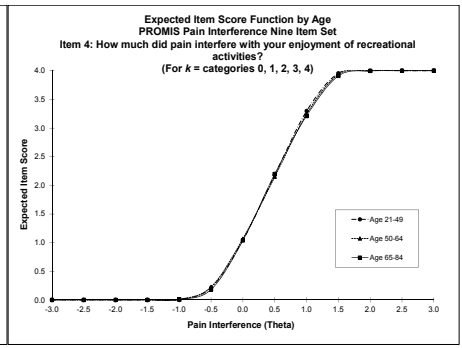


Figure 1 - cont.:
PROMIS pain interference nine item set: Expected scale score functions for gender subgroups

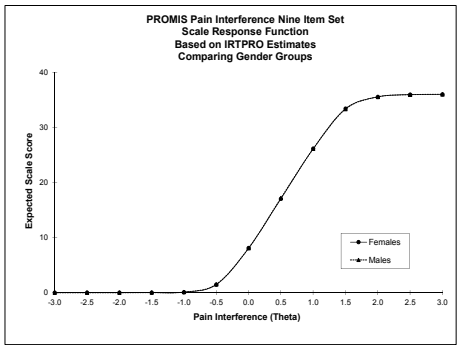
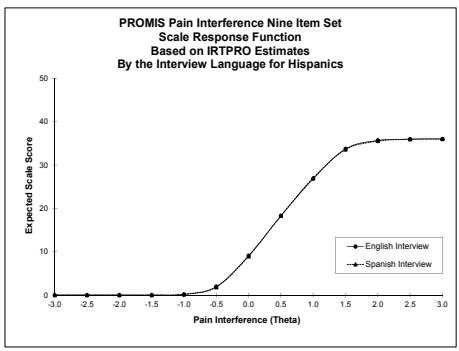
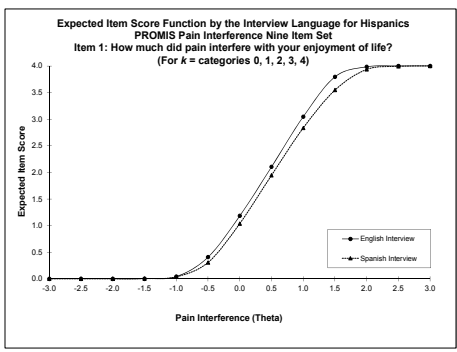


Figure 1 - cont.:
PROMIS pain interference nine item set: Expected scale score functions for interview language for Hispanics





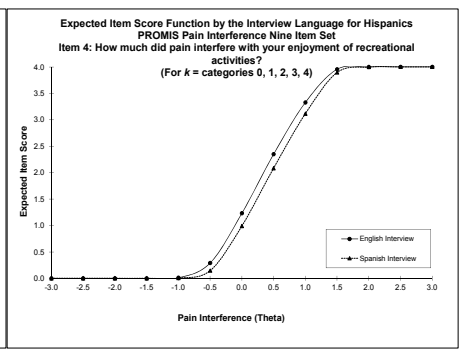


Table 8:
PROMIS pain interference nine item set: Differential item function (DIF) results. Education subgroups comparisons

Item description	IRTPRO				lordif				Magnitude (NCDIF)				Effect Size T1			
	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS
How much did pain interfere with your enjoyment of life?		NU	NU	NU*			U*	U*	0.0006	0.0016	0.0027	0.0079	-0.0028	0.0098	0.0319	0.0696
How much did pain interfere with your ability to concentrate?	NU*	NU*	NU*	NU*			U*	U*	0.0004	0.0033	0.0039	0.0008	0.0033	0.0414	0.0443	-0.0106
How much did pain interfere with your day to day activities?				U			U*	NU; U*	0.0015	0.0047	0.0047	0.0190	0.0135	0.0212	0.0154	0.0302
How much did pain interfere with your enjoyment of recreational activities?				U*			U*	NU*	0.0063	0.0085	0.0150	0.0382	0.0197	0.0115	0.0728	0.1401
How much did pain interfere with doing your tasks away from home?	Anchor item						NU; U*	U*	0.0017	0.0020	0.0009	0.0022	0.0227	0.0323	-0.0148	-0.0252
How much did pain interfere with your ability to participate in social activities?	Anchor item						U*	U*	0.0030	0.0026	0.0059	0.0045	0.0073	-0.0010	-0.0427	-0.0295
How much did pain interfere with work around the home?		U		NU		NU; U*	NU*; U*	U*	0.0018	0.0032	0.0055	0.0075	-0.0110	-0.0145	-0.0399	-0.0468
How much did pain interfere with the things you usually do for fun?				NU			NU*; U*	U*	0.0018	0.0021	0.0027	0.0086	0.0237	0.0195	0.0342	0.0684
How much did pain interfere with your enjoyment of social activities?	Item excluded															
How often did pain keep you from socializing with others?				NU; U			NU*; U*	U*	0.0006	0.0012	0.0070	0.0115	0.0142	-0.0159	-0.0492	-0.0792

All NCDIF values were smaller than the threshold (0.0960) *Asterisks indicate significance after adjustment for multiple comparisons. † Indicates value above threshold of 0.10.
NU= Non-uniform DIF involving the discrimination parameters; U=Uniform DIF involving the location parameters.
For the lordif analyses, the uniform and non-uniform DIF was determined using the likelihood ratio χ^2 test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R^2 measures of Cox & Snell, Nagelkerke, and McFadden or when using the change in Beta criterion.

Table 9:

PROMIS pain interference nine item set: Differential item function (DIF) results. Gender and age subgroups comparisons

Item description	IRTPRO			lordif			Magnitude (NCDIF)			Effect Size T1		
	Gender	Age		Gender	Age		Gender	Age		Gender	Age	
		21-49 vs. 50-64	21-49vs. 65-84		21-49 vs. 50-64	21-49vs. 65-84		21-49 vs. 50-64	21-49vs. 65-84		21-49 vs. 50-64	21-49vs. 65-84
How much did pain interfere with your enjoyment of life?	NU	NU	NU*; U	U*		NU*; U*	0.0023	0.0013	0.0055	-0.0350	-0.0030	-0.0006
How much did pain interfere with your ability to concentrate?	Anchor	NU*; U	NU*; U*		U*	NU*; U*	0.0003	0.0096	0.0391	-0.0007	0.0665	0.1328†
How much did pain interfere with your day to day activities?	Anchor	U	NU*; U	U	NU	NU*; U*	0.0003	0.0033	0.0053	-0.0081	0.0130	0.0192
How much did pain interfere with your enjoyment of recreational activities?	Anchor		NU*	U		U*	0.0009	0.0010	0.0021	-0.0163	0.0083	0.0136
How much did pain interfere with doing your tasks away from home?	U			U*		NU; U	0.0058	0.0006	0.0052	0.0553	-0.0018	-0.0164
How much did pain interfere with your ability to participate in social activities?	Anchor						0.0001	0.0025	0.0111	0.0030	-0.0270	-0.0593
How much did pain interfere with work around the home?	Anchor	Anchor item					0.0005	0.0016	0.0060	0.0154	-0.0292	-0.0506
How much did pain interfere with the things you usually do for fun?	U	Anchor item		U*			0.0023	0.0015	0.0056	-0.0325	-0.0204	-0.0290
How much did pain interfere with your enjoyment of social activities?	Item excluded											
How often did pain keep you from socializing with others?	Anchor		U		NU*; U	NU*	0.0001	0.0099	0.0138	0.0044	-0.0586	-0.0562

All NCDIF values were smaller than the threshold (0.0960) *Asterisks indicate significance after adjustment for multiple comparisons. † Indicates value above threshold of 0.10.

NU= Non-uniform DIF involving the discrimination parameters; U=Uniform DIF involving the location parameters.

For the lordif analyses, the uniform and non-uniform DIF was determined using the likelihood ratio χ^2 test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R^2 measures of Cox & Snell, Nagelkerke, and McFadden or when using the change in Beta criterion.

Gender: As shown in Table 9, no item showed gender DIF after Bonferroni correction estimated by IRTPRO; three items showed DIF with lordif, after correction for multiple comparisons: enjoyment of life, doing tasks away from home and things you do for fun. However, the magnitude of DIF was negligible. The last two items were also not among the six items hypothesized to show gender DIF (see Table 1 and Appendix Table A6).

Age: Four items showed age DIF consistently by the Wald test and OLR: enjoyment of life, ability to concentrate, day to day activities, and enjoyment of recreational activities (see Table 9). All four items showed non-uniform DIF using the primary method; all were observed to be more discriminating for the younger reference group respondents aged 21 to 49 compared with the oldest group. For one item (ability to concentrate), the differences in the *a* parameters were significant for comparisons with both oldest groups of responders (aged 50 to 64 and aged 65 to 84) after the Bonferroni adjustment. For the other three items, only the comparison between the youngest and the oldest group was significant. In addition, the item ability to concentrate showed uniform DIF; contrary to the hypothesis, the oldest respondents (aged 65 – 84) were less likely to endorse the items in the direction of pain interference compared to the youngest group (aged 21 – 49), conditional on the level of pain interference (see Appendix Table A7). This one item, ability to concentrate, showed DIF above threshold on the *T1* statistic for respondents age 65 – 84 compared to the youngest respondents age 21 – 49; however, the magnitude of DIF was low and none of the NCDIF statistics were above threshold. An additional item showed DIF estimated by the OLR method for the comparison of the youngest to the oldest age group, socializing with others. Older respondents were hypothesized to report higher levels of pain interference on all four items that evidenced significant DIF; however, non-uniform DIF was detected for these items.

Language: DIF analysis was performed for Hispanics only, contrasting those interviewed in English with those interviewed in Spanish. Spanish speakers were hypothesized to experience less pain interference on one item, enjoyment of life (see Table 1). The DIF findings confirmed the hypothesis. The item showed uniform DIF in favor of the Spanish speakers (less pain interference) as did an additional item, enjoyment of recreational activities (see Table 10 and Appendix Table A8). Both items showed DIF above threshold on the *T1* statistic; however, the magnitude was small, and all NCDIF statistics were below the threshold.

Sensitivity analyses for anchor item selection

Sensitivity analyses were performed to evaluate the effect of less than optimal numbers of anchor items. The number of anchor items was sufficient for the analyses of gender and language of interview. However, only two items were identified as anchors in the analyses of race/ethnicity: enjoyment of life and work around the home. When the data were reanalyzed with four anchors, adding the items day to day activities and doing your tasks away from home; the same items as shown in Table 7 were identified with DIF after the Bonferroni adjustment. There were three changes: uniform DIF for the item enjoyment of recreational activities became significant after Bonferroni correction for Black and Hispanic respondents compared to the non-Hispanic White respondents. The

item ability to participate in social activities changed from significant uniform DIF after Bonferroni correction to DIF before the correction for Hispanics (data not shown).

Two items were identified as anchors in the education DIF analysis: doing your tasks away from home and ability to participate in social activities (see Table 8). The additional two anchors in the sensitivity reanalysis were: enjoyment of life and things you usually do for fun. Overall, one additional item showed uniform DIF after the Bonferroni

Table 10:
PROMIS pain interference item set: Differential item function (DIF) results. Language subgroups comparison, English (n = 704) vs. Spanish (n = 334) interview, for Hispanics only

Item description	IRTPRO	lordif	Magnitude (NCDIF)	Effect Size T1
How much did pain interfere with your enjoyment of life?	U*	U*	0.0199	0.1153†
How much did pain interfere with your ability to concentrate?	Anchor item	U*	0.0082	-0.0700
How much did pain interfere with your day to day activities?	U	NU*; U*	0.0099	0.0070
How much did pain interfere with your enjoyment of recreational activities?	U*	NU; U*	0.0367	0.1539†
How much did pain interfere with doing your tasks away from home?	Anchor item	NU*; U*	0.0100	-0.0706
How much did pain interfere with your ability to participate in social activities?	Anchor item	NU; U*	0.0009	-0.0217
How much did pain interfere with work around the home?	Anchor item	NU; U*	0.0063	-0.0626
How much did pain interfere with the things you usually do for fun?	Anchor item	NU*; U*	0.0035	0.0083
How much did pain interfere with your enjoyment of social activities?	Item excluded			
How often did pain keep you from socializing with others?	Anchor item	U*	0.0064	-0.0298

All NCDIF values were smaller than the threshold (0.0960) *Asterisks indicate significance after adjustment for multiple comparisons. † Indicates value above threshold of 0.10.
NU = Non-uniform DIF involving the discrimination parameters; U = Uniform DIF involving the location parameters.
For the lordif analyses, the uniform and non-uniform DIF was determined using the likelihood ratio χ^2 test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R^2 measures of Cox & Snell, Nagelkerke, and McFadden or when using the change in Beta criterion.

correction: socializing with others for respondents with less than high school education compared to those with a graduate degree. The change in the DIF designation was for the item ability to concentrate for respondents with a college degree and some college vs. those with a graduate degree. The change was from non-uniform DIF after Bonferroni adjustment to non-uniform DIF prior to adjustment (data not shown).

No changes were observed in the age sensitivity re-analysis. Two anchor items were originally identified: work around the home and things you usually do for fun. The additional anchors were: enjoyment of recreational activities and doing your tasks away from home (data not shown).

Aggregate DIF Impact

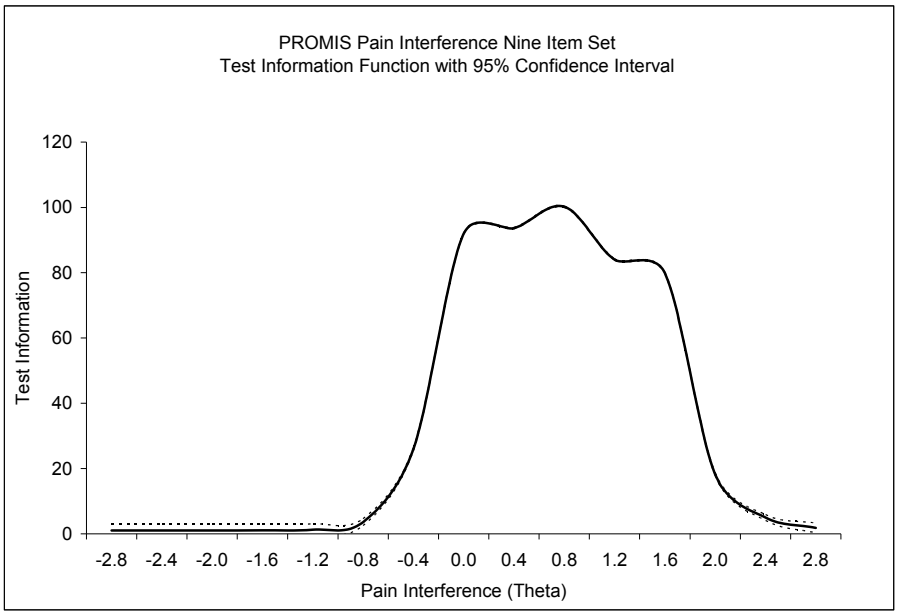
As shown in Figure 1, there was no evident scale level impact. All group curves were overlapping for all comparisons.

Individual DIF Impact

Individual impact analysis was performed by comparing θs estimated accounting for and not accounting for DIF for race/ethnicity, education, age, and language group comparisons. The evaluation was not performed for the gender groups because no significant DIF was observed after Bonferroni correction. All correlations between the two sets of θs were 1.0 and the differences between both estimates were minor, all within 0.05 standard deviations. Theta estimates were slightly higher overall when estimates accounted for DIF. For non-Hispanic Whites, DIF-adjusted θs were higher compared with unadjusted θs, Black respondents were almost evenly split (49 % with higher DIF adjusted θs); however, for the majority of Hispanics (85 %) and all Asians/Pacific Islanders the unadjusted θs were higher compared to the adjusted θs. There was no clear pattern of differences among the education groups; 65 % of respondents overall evidenced adjusted θs higher than unadjusted θs. Among the age groups, 98 % of respondents age 50 to 64 evidenced unadjusted θs higher than adjusted θs as contrasted with respondents aged 21 to 49 (49 %) and aged 65 to 84 (47 %). English speakers evidenced higher initial θs (68 %) than Spanish speakers (32 %).

Using an arbitrary cutoff point of $\theta \geq 1.0$ to classify respondents with acute pain interference, the highest number of changes were for the education groups analyses. There were 231 respondents (4% of the total sample) who changed from the designation of no acute pain interference to acute interference after the DIF adjustment. There was a relationship between education and designation change: the lower the education, the higher the percent of respondents who changed designations from without acute pain interference to pain interference after DIF adjustment (1.6% for the graduate degree respondents to 7.2% for respondents with less than high school). There was no change in the designations for race/ethnic subgroups. Among the age groups there were a small number of changes for respondents aged 65 to 84 (22 or 1.0%).

Figure 2:
PROMIS pain interference nine item set: Scale (test) information function (IRTPRO; Total sample)



Information

The item-level and scale information functions were examined for the total sample (see Figure 2 and Appendix Figure A2).The scale and the individual item information functions were very high especially along the θ continuum from 0 to 1.6 reaching 100.2 for the scale information at $\theta = 0.8$. Little or no information was provided below θ of -0.6 and above θ of 2.1 for the total scale. The most informative items were "How much did pain interfere with the things you usually do for fun?" (peak information of 16.33 at $\theta = 0$); pain interference in ability to participate in social activities (peak of 14.88 at $\theta = 0.8$); and pain interference with work around the home (peak of 14.70 at $\theta = 0.8$). The least informative items overall were: pain interference with ability to concentrate (peak of 5.06 at $\theta = 0.8$) and "How often did pain keep you from socializing with others?" (peak of 5.51 at $\theta = 0.4$); however these items provided some information at the high levels of the pain interference continuum ($\theta = 2.0$ to 2.8) where other items were not informative.

Discussion

Similar to the study by Amtmann and colleagues (2010), age DIF was identified for the items related to concentration and gender DIF for the item related to enjoyment of life, but only for one method and the result was of low magnitude. The item related to con-

centration was hypothesized to show uniform DIF in the direction that those with lower education would report more interference; however, only non-uniform DIF was observed. Although there were no hypotheses related to race/ethnicity, several items evidenced DIF, two of somewhat higher magnitude for Hispanic and Asians/Pacific Islanders in contrast to non-Hispanic Whites: ability to concentrate and engagement in recreational activities. Several items evidenced DIF for education; however, none had accompanying hypotheses related to DIF, except for one item, ability to concentrate. Because non-uniform DIF was observed for that item, the hypothesis of uniform DIF was not confirmatory. No items evidenced DIF of high magnitude.

Limitations: The fit for the unidimensional model was somewhat lower for Hispanics interviewed in English than for the other groups; however, the CFI values were very high (> 0.99) across all groups. Although the local dependencies were high overall, within groups they were more reasonable, particularly given the smaller sample sizes. The LD statistic appears to be affected by sample size. Nonetheless, one item pair was particularly problematic: interference with things you do for fun and enjoyment of social activities. Because of the high local dependency values, the discrimination parameter estimates were also inflated. Given the presence of another item related in content: interference with enjoyment of recreational activities; it was decided to drop the enjoyment of social activities item.

Another limitation was the inability to identify adequate numbers of anchor items. Sensitivity analyses identified some changes in results when the anchor sets were increased to include up to two additional items with some low magnitude of DIF. These changes were usually in the direction of identification of additional DIF. Because DIF in anchor items can result in false DIF detection, it is unknown if the sensitivity analyses were thus affected. A final limitation is that not all of the short form items were examined; only one short form included all items (6b); the others were missing one or two items, in part due to the concurrent selection of items for this survey and for the short forms developed by the PROMIS investigators.

Summary: In summary, the items can be recommended for use with persons with cancer and those in palliative care because the measure performed well in terms of reliability (estimates were generally 0.90 and above) and precision. Relatively high information was provided in the mid to high range of pain interference. In general the magnitude of DIF was very small, and the aggregate impact not measureable. However, some individual impact was observed for education, primarily among those with lower education. Finally, one item was removed from the analyses due to assumption violations. Thus, the nine pain interference short form items and associated scales examined in these analyses can be recommended for use among ethnically diverse groups, with the caveat that there may be an impact of DIF on the scores of some individuals with lower education.

Acknowledgements

Partial funding for the analyses was provided by the National Institute on Aging (NIA)-funded Mt. Sinai Pepper Center, P30AG028741 (PI: Siu), the National Institute of Ar-

thritis & Musculoskeletal & Skin Diseases, U01AR057971 (PI: Potosky, Moinpour) and the NIA Edward R. Roybal Center, P30AG022845 (PI: Reid, Pillemer, Wethington).The authors thank Stephanie Silver, MPH for editorial assistance in the preparation of this manuscript.

References

Amtmann, D., Cook, K., Jensen, M. P., Chen, W-H., Choi, S., Revicki, D., … Lai, J-S. (2010). Development of a PROMIS item bank to measure pain interference. *Pain, 150*, 173-182. doi:10.1016/j.pain.2010.04.025

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438.doi: 10.1016/j.jrp.2014.07.001.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. doi: 10.1037/0033-2909.107.2.238

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Breivik, H., Collett, B., Ventafridda, V., Cohen, R., & Gallacher, D. (2006). Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. *European Journal of Pain, 10*, 287-333. doi:10.1016/j.ejpain.2005.06.009

Cai, L., Thissen, D., & du Toit, S. H.C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]. Chicago, IL: Scientific Software International, Inc.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., … Rose, M., on behalf of the PROMIS Cooperative Group. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3-S11. doi:10.1097/01.mlr.0000258615.42478.55

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289. doi:10.3102/10769986022003265

Chen, W-H., Revicki, D., Lai, J-S., Cook, K. F., & Amtmann, D. (2009). Linking pain items from two studies using item response theory. *Journal of Pain and Symptom Management, 38*, 615-628. doi:10.1016/j.jpainsymman.2008.11.016

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*, 1-30. doi:10.18637/jss.v039.i08

Cleeland, C. S. (1989). Measurement of pain by subjective report. In C. R. Chapman & J. D. Loeser (Eds.), *Advances in pain research and therapy, volume 12: Issues in pain measurement* (pp. 393-403). New York: Raven Press.

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*, 447-460. doi:10.1007/s11136-009-9464-4

Cox, D. R., & Snell, E. J. (1989).*The analysis of binary data* (2nd Ed.). London: Chapman and Hall.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Medical Care, 44*(11 Suppl 3), S115-S123. doi:10.1097/01.mlr.0000245183.28384.ed

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241-256. doi:10.1002/sim.1713

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. on behalf of the PROMIS cooperative group. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(5 Suppl 1), S12-S21. doi:10.1097/01.mlr.0000254567.79743.e2

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement, 23*, 309-326.doi:10.1177/01466219922031437

Jensen, R. E., Moinpour, C. M., Keegan, T. H. M., Cress, R. D., Wu, X.-C., Paddock, L. A., . . . Potosky, A. L. (2016). The Measuring Your Health Study: Leveraging community-based cancer registry recruitment to establish a large, diverse cohort of cancer survivors for analyses of measurement equivalence and validity of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short form items. *Psychological Test and Assessment Modeling, 58*(1), 99-117.

Johannes, C. B., Le, T. K., Zhou, X., Johnston, J. A., & Dworkin, R. H. (2010). The prevalence of chronic pain in United States adults: Results of an internet-based survey. *The Journal of Pain, 11*, 1230-1239. doi:10.1016/j.jpain.2010.07.002

Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*, 93-116. doi:10.1111/j.1745-3984.2007.00029.x

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling, 58(1)*, 79-98.

Kopec, J. A., Badii, M., McKenna, M., Lima, V. D., Sayre, E. C., & Dvorak, M. (2008). Computerized adaptive testing in back pain: Validation of the CAT-5D-QOL. *Spine, 33*, 1384-1390. doi:10.1097/BRS.0b013e3181732a3b

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: L. Erlbaum Associates.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

Meade, A. W., Johnson, E. C., & Bradley, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*, 568-592. doi:10.1037/0021-9010.93.3.568

Moulin, D. E., Clark, A. J., Speechley, M., & Morley-Forester, P. K. (2002). Chronic pain in Canada - Prevalence, treatment, impact and the role of opioid analgesia. *Pain Research and Management, 7*, 179-184. doi:10.1155/2002/323085

Mukherjee, S., Gibbons, L. E., Kristiansson, E., & Crane, P. K. (2013). Extension of an iterative hybrid ordinal logistic regression/item response theory approach to detect and account for differential item functioning in longitudinal data. *Psychological Test and Assessment Modeling, 55*, 127-147.

Muthén, L. K., & Muthén, B. O. (2011). *M-PLUS users guide* (6th ed.). Los Angeles, California: Muthén and Muthén.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691-692. doi:10.1093/biomet/78.3.691

Orlando-Edelen, M., & Saliba, D. (2010). Correspondence of verbal descriptor and numeric rating scales for pain intensity: An item response theory calibration. *Journal of Gerontology, Medical Sciences, 65*, 778-785. doi:10.1093/gerona/glp215

Orlando-Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Applications to the Mini-Mental State Examination. *Medical Care, 44*(11 Suppl 3), S134-S142. doi:10.1097/01.mlr.0000245251.83359.8c

Oshima, T. C., Kushubar, S., Scott, J. C., & Raju, N. S. (2009). *DFIT for Window User's Manual: Differential functioning of items and tests.* St. Paul, MN: Assessment Systems Corporation.

Raju, N. S. (1999). DFITP5: A Fortran program for calculating dichotomous DIF/DTF [Computer program]. Chicago: Illinois Institute of Technology.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M., L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the DFIT framework. *Applied Measurement in Education, 33*, 133-147. doi:10.1177/0146621608319514

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368. doi:10.1177/014662169501900405.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., … Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the Patient-Reported Outcome Measurement Information System (PROMIS). *Medical Care, 45*(5 Suppl 1), S22-S31. doi:10.1097/01.mlr.0000250483.85507.04

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research,47*, 667-696. doi:10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bi-factor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544-559. doi:10.1080/00223891.2010.496477

Rizopoulus, D. (2009). ltm: Latent Trait Models under IRT. Retrieved from: http://cran.r project.org/web/packages/ltm/index.html.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100-114.

Schmid, L., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61. doi:10.1007/BF02289209

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120. doi:10.1007/s11336-008-9101-0

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370. doi:10.1111/j.1745-3984.1990.tb00754.x

Teresi, J. A., & Jones, R. N. (2016). Methodological issues in examining measurement equivalence in patient reported outcomes measures: Methods overview to the two-part series, "Measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) short form measures". *Psychological Test and Assessment Modeling*, *58*(1)*,* 37-78.

Wainer, H. (1993). Model-based standardization measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 123-135). Hillsdale, NJ: Lawrence Erlbaum, Inc.

van Nispen tot Pannerden, S. C., Candel, M. J. J. M., Zwakhalen, S. M. G., Hamers, J. P. H., Curfs, L. M. G., & Berger, M. P. F. (2009). An item response theory-based assessment of pain assessment checklist for seniors with limited ability to communicate (PACSLAC). *The Journal of Pain, 10*, 844-853. doi:10.1016/j.jpain.2009.02.007

Walton, D. M., Wideman, T. H., & Sullivan, M. J. L. (2013). A Rasch analysis of the pain catastrophizing scale supports its use as an interval-level measure. *Clinical Pain, 29*, 499-505. doi:10.1097/AJP.0b013e318269569c

Waterman, C., Victor, T. W., Jensen M. P., Gould, E. M., Gammaitoni, A. R., & Galer, B. S. (2010). The assessment of pain quality: An item response theory analysis. *The Journal of Pain,11*, 273-279. doi:10.1016/j.jpain.2009.07.014

Zumbo, B. D. (1999). *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. Retrieved from: http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.

Appendix

Table A1:
PROMIS pain interference item set. Model fit statistics: Comparative fit index (CFI) from the confirmatory and bi-factor models (10 item set) and graded response models fit from IRTPRO (10 and 9 item sets) for the total sample and demographic subgroups

Sample	CFA CFI (MPLUS)	IRT Model RMSEA (IRTPRO)	
		10 item set	9 item set
Total Sample (CFA)	0.998	0.05	0.05
Random First Half Sample (CFA)	0.998	N/A	N/A
Random Second Half Sample (Bi-factor CFA)	0.999	N/A	N/A
Female	0.998	0.06	0.05
Male	0.998	0.06	0.05
Age 21 to 49 years	0.998	0.08	0.07
Age 50 to 64 years	0.998	0.06	0.05
Age 65 to 84 years	0.997	0.05	0.05
Non-Hispanic Whites	0.998	0.05	0.05
Blacks	0.997	0.07	0.07
Hispanics	0.997	0.08	0.07
Asian/Pacific Islanders	0.998	0.06	0.06
Less Than High School	0.997	0.08	0.07
High School Graduate	0.998	0.08	0.07
Some College	0.998	0.06	0.05
College Graduate	0.998	0.07	0.06
Graduate Degree	0.998	0.05	0.05
Hispanics Interviewed in English	0.997	0.11	0.10
Hispanics Interviewed in Spanish	0.998	0.07	0.05

Table A2:
PROMIS pain interference ten item set: Discrimination '*a*' parameter estimates. IRTPRO individual sub-sample runs

Item	Item description	Total Sample	Race/Ethnicity				Education					Age Groups			Gender		Hispanics Only Interview Language	
			NH Whites	NH Black	Hisp.	NH API	No HS	HS	Some Coll.	College	Grad.	21-49	50-64	65-84	Female	Male	Engl.	Span.
1	How much did pain interfere with your enjoyment of life?	4.94	5.23	4.86	4.75	5.35	4.37	4.92	5.30	4.91	6.83	5.80	6.03	4.41	5.15	5.50	5.11	4.32
2	How much did pain interfere with your ability to concentrate?	4.00	3.94	3.76	4.17	5.07	3.66	3.73	4.14	3.94	6.13	4.81	4.80	3.58	4.01	4.74	4.33	3.91
3	How much did pain interfere with your day to day activities?	6.15	6.72	5.83	6.01	6.47	5.84	6.23	6.37	5.93	7.80	7.02	7.62	5.53	6.15	7.29	6.36	5.61
4	How much did pain interfere with your enjoyment of recreational activities?	6.11	6.43	6.40	6.58	6.16	5.86	6.92	6.37	6.13	6.65	6.64	7.35	5.73	6.07	7.38	6.90	6.56
5	How much did pain interfere with doing your tasks away from home?	7.08	7.28	7.13	6.47	7.95	6.88	7.28	7.06	6.74	9.06	7.03	9.03	6.79	7.28	8.19	6.36	6.78
6	How much did pain interfere with your ability to participate in social activities?	9.07	9.63	8.70	8.61	9.69	8.30	9.28	9.62	8.60	10.34	8.47	10.98	9.14	8.68	11.33	8.84	8.00
7	How much did pain interfere with work around the home?	7.49	8.17	6.78	7.27	8.42	6.58	7.29	8.01	7.50	9.56	7.47	8.67	7.55	7.64	8.52	7.55	6.79
8	How much did pain interfere with the things you usually do for fun?	8.91	9.13	8.48	8.96	10.43	7.90	10.18	9.01	8.69	10.72	8.02	10.75	9.37	9.30	10.13	8.66	9.91
9	How much did pain interfere with your enjoyment of social activities?	9.63	10.71	8.48	9.09	10.89	8.32	9.76	10.01	10.62	12.19	8.56	10.93	10.91	9.52	11.76	9.66	8.03
10	How often did pain keep you from socializing with others?	4.37	4.87	4.45	4.11	4.46	3.73	4.32	4.84	4.37	5.57	4.02	5.24	4.57	4.46	5.07	4.48	3.44

Table A3:
PROMIS pain interference nine item set: Discrimination '*a*' parameter estimates. IRTPRO individual sub-sample runs

Item	Item description	Total Sample	Race/Ethnicity				Education					Age Groups			Gender		Hispanics Only Interview Language	
			NH Whites	NH Black	Hisp.	NH API	No HS	HS	Some Coll.	College	Grad.	21-49	50-64	65-84	Female	Male	Engl.	Span.
1	How much did pain interfere with your enjoyment of life?	5.25	5.55	5.12	5.02	5.40	4.65	5.21	5.58	5.31	6.39	6.39	5.68	4.67	5.52	5.16	5.35	4.63
2	How much did pain interfere with your ability to concentrate?	4.16	4.06	3.89	4.37	5.06	3.85	3.85	4.28	4.09	5.55	5.17	4.35	3.70	4.17	4.33	4.53	4.12
3	How much did pain interfere with your day to day activities?	6.76	7.48	6.30	6.59	6.63	6.42	6.75	6.92	6.72	7.48	8.01	7.40	6.01	6.79	7.02	7.06	5.98
4	How much did pain interfere with your enjoyment of recreational activities?	6.47	6.82	6.64	6.92	6.22	6.09	7.45	6.69	6.56	6.21	7.36	6.82	6.05	6.41	6.93	7.28	6.85
5	How much did pain interfere with doing your tasks away from home?	7.28	7.50	7.40	6.56	7.76	7.07	7.32	7.30	6.81	8.56	7.45	8.15	6.90	7.51	7.41	6.42	6.94
6	How much did pain interfere with your ability to participate in social activities?	8.21	8.53	8.13	7.65	8.59	7.68	8.37	8.67	7.36	8.02	7.82	8.80	8.17	7.84	9.08	7.69	7.39
7	How much did pain interfere with work around the home?	7.73	8.60	6.92	7.08	8.58	6.62	7.50	8.29	7.89	9.05	8.01	7.97	7.83	7.94	7.86	7.24	6.80
8	How much did pain interfere with the things you usually do for fun?	8.03	8.50	7.48	7.65	8.90	6.96	9.18	7.97	7.65	9.16	7.63	8.57	8.29	8.20	8.22	7.28	8.61
10	How often did pain keep you from socializing with others?	4.34	4.84	4.45	3.96	4.17	3.73	4.28	4.74	4.24	4.92	4.09	4.60	4.52	4.42	4.41	4.32	3.34

Table A4:
PROMIS pain interference nine item set: IRT item parameters and DIF statistics for the race/ethnic groups; non-Hispanic Whites are the reference group

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a* DIF*	*b* DIF*
How much did pain interfere with your enjoyment of life?	Non-Hispanic White	5.34 (0.13)	-0.09 (0.02)	0.54 (0.01)	1.01 (0.01)	1.59 (0.02)	NS, Anchor item	
	Black							
	Hispanic							
	Asian/Pacific Islander							
How much did pain interfere with your ability to concentrate?	Non-Hispanic White	4.05 (0.15)	0.37 (0.02)	0.88 (0.02)	1.41 (0.03)	1.96 (0.05)		
	Black	3.92 (0.19)	0.41 (0.03)	0.88 (0.03)	1.37 (0.04)	1.91 (0.06)	0.4 (0.520)	3.3 (0.503)
	Hispanic	4.44 (0.21)	0.25 (0.03)	0.75 (0.03)	1.27 (0.03)	1.85 (0.05)	2.2 (0.139)	***20.5 (<0.001)***
	Asian/Pacific Islander	5.25 (0.29)	0.12 (0.03)	0.70 (0.03)	1.10 (0.03)	1.72 (0.06)	***12.2 (0.001)***	***69.6 (<0.001)***
How much did pain interfere with your day to day activities?	Non-Hispanic White	7.42 (0.30)	-0.03 (0.02)	0.56 (0.02)	0.98 (0.02)	1.55 (0.03)		
	Black	6.33 (0.31)	-0.03 (0.03)	0.59 (0.02)	1.01 (0.02)	1.57 (0.03)	6.7 (0.010)	1.1 (0.902)
	Hispanic	6.68 (0.35)	-0.04 (0.03)	0.57 (0.02)	0.95 (0.02)	1.50 (0.03)	2.4 (0.125)	4.3 (0.372)
	Asian/Pacific Islander	6.91 (0.41)	-0.07 (0.03)	0.56 (0.02)	0.95 (0.03)	1.56 (0.04)	1.3 (0.252)	2.6 (0.633)
How much did pain interfere with your enjoyment of recreational activities?	Non-Hispanic White	6.78 (0.25)	-0.07 (0.02)	0.45 (0.02)	0.82 (0.02)	1.31 (0.02)		
	Black	6.66 (0.33)	-0.02 (0.03)	0.52 (0.02)	0.87 (0.02)	1.42 (0.03)	0.2 (0.651)	13.4 (0.010)
	Hispanic	7.03 (0.37)	-0.02 (0.03)	0.48 (0.02)	0.89 (0.02)	1.40 (0.03)	0.3 (0.594)	***14.4 (0.006)***
	Asian/Pacific Islander	6.43 (0.36)	-0.12 (0.03)	0.53 (0.02)	0.90 (0.03)	1.44 (0.04)	0.8 (0.376)	***21.2 (<0.001)***
How much did pain interfere with doing your tasks away from home (e.g., getting groceries, running errands)?	Non-Hispanic White	7.48 (0.32)	0.17 (0.02)	0.65 (0.02)	1.06 (0.02)	1.49 (0.03)		
	Black	7.50 (0.42)	0.14 (0.02)	0.61 (0.02)	1.02 (0.02)	1.52 (0.03)	<0.1 (0.939)	7.2 (0.127)
	Hispanic	6.70 (0.35)	0.11 (0.02)	0.58 (0.02)	0.99 (0.02)	1.55 (0.03)	2.5 (0.112)	12.4 (0.015)
	Asian/Pacific Islander	7.92 (0.53)	0.05 (0.02)	0.57 (0.02)	0.97 (0.03)	1.46 (0.04)	0.4 (0.524)	***20.0 (0.001)***
How much did pain interfere with your ability to participate in social activities?	Non-Hispanic White	8.61 (0.45)	0.23 (0.02)	0.65 (0.02)	1.08 (0.02)	1.47 (0.02)		
	Black	8.18 (0.49)	0.14 (0.02)	0.62 (0.02)	1.01 (0.02)	1.56 (0.03)	0.6 (0.441)	***28.0 (<0.001)***
	Hispanic	7.80 (0.47)	0.13 (0.02)	0.61 (0.02)	1.02 (0.02)	1.52 (0.03)	1.6 (0.210)	***16.2 (0.003)***
	Asian/Pacific Islander	8.80 (0.72)	0.12 (0.02)	0.60 (0.02)	0.97 (0.02)	1.48 (0.04)	<0.1 (0.876)	***25.3 (<0.001)***
How much did pain interfere with work around the home?	Non-Hispanic White	7.86 (0.22)	-0.03 (0.02)	0.53 (0.01)	0.94 (0.01)	1.49 (0.02)	NS, Anchor item	
	Black							
	Hispanic							
	Asian/Pacific Islander							

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a* DIF*	*b* DIF*
How much did pain interfere with the things you usually do for fun?	Non-Hispanic White	8.19 (0.24)	0.02 (0.01)	0.53 (0.01)	0.93 (0.01)	1.40 (0.02)	DIF not significant	
	Black							
	Hispanic							
	Asian/Pacific Islander							
How much did pain interfere with your enjoyment of social activities?	Non-Hispanic White	Item not included in the analysis						
	Black							
	Hispanic							
	Asian/Pacific Islander							
How often did pain keep you from socializing with others?	Non-Hispanic White	4.81 (0.18)	0.27 (0.02)	0.67 (0.02)	1.20 (0.02)	1.93 (0.05)		
	Black	4.47 (0.21)	0.13 (0.03)	0.57 (0.02)	1.21 (0.03)	1.99 (0.06)	1.9 (0.169)	***21.3 (<0.001)***
	Hispanic	4.05 (0.19)	0.12 (0.03)	0.52 (0.03)	1.20 (0.03)	1.88 (0.06)	***8.3 (0.004)***	***29.5 (<0.001)***
	Asian/Pacific Islander	4.31 (0.23)	0.05 (0.03)	0.54 (0.03)	1.21 (0.04)	1.80 (0.07)	3.2 (0.073)	***44.5 (<0.001)***

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters.

* Bolded entries indicate items that evidence DIF after correction for multiple comparisons; "NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Table A5:
PROMIS pain interference nine item set: IRT item parameters and DIF statistics for the education groups (reference group is graduate degree)

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a* DIF*	*b* DIF*
How much did pain interfere with your enjoyment of life?	Less than HS	5.09 (0.26)	-0.24 (0.03)	0.38 (0.02)	0.80 (0.03)	1.36 (0.04)	***9.2 (0.002)***	4.2 (0.378)
	High School	5.27 (0.28)	-0.27 (0.03)	0.32 (0.02)	0.78 (0.03)	1.39 (0.04)	7.2 (0.007)	1.6 (0.801)
	Some College	5.67 (0.24)	-0.34 (0.03)	0.27 (0.02)	0.75 (0.02)	1.31 (0.03)	4.1 (0.042)	4.1 (0.390)
	College Degree	5.61 (0.34)	-0.39 (0.04)	0.27 (0.03)	0.76 (0.03)	1.29 (0.05)	3.8 (0.050)	6.4 (0.173)
	Graduate Degree	6.76 (0.48)	-0.29 (0.03)	0.33 (0.03)	0.73 (0.04)	1.35 (0.07)		
How much did pain interfere with your ability to concentrate?	Less than HS	4.23 (0.22)	0.08 (0.03)	0.58 (0.03)	1.00 (0.03)	1.61 (0.05)	***9.7 (0.002)***	4.7 (0.321)
	High School	3.92 (0.21)	0.14 (0.03)	0.60 (0.03)	1.12 (0.04)	1.68 (0.06)	***14.5 (<0.001)***	6.0 (0.201)
	Some College	4.35 (0.18)	0.07 (0.02)	0.58 (0.02)	1.10 (0.03)	1.64 (0.05)	***8.9 (0.003)***	1.4 (0.844)
	College Degree	4.31 (0.26)	0.04 (0.03)	0.56 (0.03)	1.05 0.05	1.53 (0.08)	***8.0 (0.005)***	2.9 (0.580)
	Graduate Degree	5.75 (0.44)	0.05 (0.03)	0.58 (0.04)	1.04 0.05	1.57 (0.10)		
How much did pain interfere with your day to day activities?	Less than HS	7.03 (0.40)	-0.17 (0.03)	0.40 (0.02)	0.73 (0.02)	1.23 (0.03)	1.4 (0.246)	10.6 (0.031)
	High School	6.84 (0.39)	-0.25 (0.03)	0.34 (0.02)	0.75 (0.02)	1.34 (0.04)	2.1 (0.151)	2.6 (0.631)
	Some College	7.05 (0.32)	-0.28 (0.02)	0.30 (0.02)	0.75 (0.02)	1.30 (0.03)	1.4 (0.231)	6.5 (0.163)
	College Degree	7.06 (0.46)	-0.32 (0.03)	0.31 (0.02)	0.72 (0.03)	1.32 (0.05)	1.1 (0.285)	3.4 (0.496)
	Graduate Degree	7.87 (0.60)	-0.26 (0.03)	0.36 (0.03)	0.71 (0.03)	1.39 (0.07)		
How much did pain interfere with your enjoyment of recreational activities?	Less than HS	6.65 (0.36)	-0.13 (0.03)	0.34 (0.02)	0.69 (0.02)	1.18 (0.03)	<0.1 (0.898)	***18.8 (0.001)***
	High School	7.53 (0.44)	-0.26 (0.03)	0.30 (0.02)	0.66 (0.02)	1.18 (0.03)	2.3 (0.133)	6.0 (0.198)
	Some College	6.82 (0.29)	-0.33 (0.03)	0.22 (0.02)	0.59 (0.02)	1.06 (0.02)	0.2 (0.658)	3.3 (0.517)
	College Degree	6.89 (0.43)	-0.33 (0.03)	0.21 (0.02)	0.56 (0.03)	1.14 (0.04)	0.3 (0.612)	5.8 (0.220)
	Graduate Degree	6.58 (0.45)	-0.31 (0.03)	0.23 (0.03)	0.66 (0.04)	1.12 (0.05)		
How much did pain interfere with doing your tasks away from home (e.g., getting groceries, running errands)?	Less than HS High School Some College College Degree Graduate Degree	7.57 (0.21)	-0.09 (0.01)	0.38 (0.01)	0.78 (0.01)	1.26 (0.02)	NS, Anchor item	
How much did pain interfere with your ability to participate in social activities?	Less than HS High School Some College College Degree Graduate Degree	8.42 (0.26)	-0.05 (0.01)	0.40 (0.01)	0.80 (0.01)	1.26 (0.02)	NS, Anchor item	

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a* DIF*	*b* DIF*
How much did pain interfere with work around the home?	Less than HS	7.22 (0.41)	-0.22 (0.03)	0.31 (0.02)	0.67 (0.02)	1.22 (0.03)	5.1 (0.023)	4.0 (0.413)
	High School	7.62 (0.44)	-0.24 (0.03)	0.30 (0.02)	0.69 (0.02)	1.26 (0.03)	3.3 (0.070)	5.4 (0.255)
	Some College	8.37 (0.41)	-0.29 (0.02)	0.28 (0.02)	0.71 (0.02)	1.26 (0.03)	1.1 (0.302)	13.0 (0.011)
	College Degree	8.46 (0.60)	-0.24 (0.03)	0.27 (0.02)	0.71 (0.03)	1.23 (0.04)	0.7 (0.395)	7.0 (0.133)
	Graduate Degree	9.34 (0.84)	-0.17 (0.03)	0.37 (0.03)	0.73 (0.03)	1.17 (0.05)		
How much did pain interfere with the things you usually do for fun?	Less than HS	7.61 (0.45)	-0.12 (0.03)	0.34 (0.02)	0.71 (0.02)	1.16 (0.03)	3.9 (0.050)	6.1 (0.190)
	High School	9.26 (0.67)	-0.21 (0.03)	0.31 (0.02)	0.72 (0.02)	1.19 (0.03)	0.1 (0.806)	4.9 (0.297)
	Some College	8.10 (0.37)	-0.23 (0.02)	0.26 (0.02)	0.67 (0.02)	1.15 (0.02)	2.3 (0.133)	4.2 (0.380)
	College Degree	8.17 (0.55)	-0.22 (0.03)	0.30 (0.02)	0.66 (0.03)	1.12 (0.04)	1.7 (0.188)	0.8 (0.941)
	Graduate Degree	9.53 (0.87)	-0.19 (0.03)	0.33 (0.03)	0.67 (0.03)	1.10 (0.05)		
How much did pain interfere with your enjoyment of social activities?		Item not included in the analysis						
How often did pain keep you from socializing with others?	Less than HS	4.11 (0.21)	-0.05 (0.03)	0.29 (0.03)	0.95 (0.03)	1.64 (0.05)	5.8 (0.016)	10.0 (0.040)
	High School	4.34 (0.22)	-0.12 (0.03)	0.36 (0.03)	1.02 (0.03)	1.73 (0.06)	3.4 (0.067)	9.2 (0.057)
	Some College	4.79 (0.19)	-0.05 (0.02)	0.37 (0.02)	0.92 (0.02)	1.66 (0.05)	0.7 (0.400)	5.3 (0.263)
	College Degree	4.51 (0.26)	-0.03 (0.03)	0.40 (0.03)	1.00 (0.04)	1.70 (0.09)	1.9 (0.168)	1.7 (0.789)
	Graduate Degree	5.14 (0.37)	0.02 (0.03)	0.43 (0.04)	0.97 0.05	1.55 (0.10)		

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters.

* Bolded entries indicate items that evidence DIF after correction for multiple comparisons; "NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Table A6:
PROMIS pain interference nine item set: IRT item parameters and DIF statistics for the gender groups, females are the reference group

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a* DIF*	*b* DIF*
How much did pain interfere with your enjoyment of life?	Males	4.96 (0.17)	-0.38 (0.02)	0.27 (0.02)	0.75 (0.02)	1.35 (0.03)	5.6(0.018)	6.0(0.197)
	Females	5.51 (0.15)	-0.33 (0.02)	0.31 (0.01)	0.78 (0.02)	1.36 (0.03)		
How much did pain interfere with your ability to concentrate?	Males	4.17 (0.10)	0.05 (0.01)	0.57 (0.02)	1.07 (0.02)	1.65 (0.03)	NS, Anchor item	
	Females							
How much did pain interfere with your day to day activities?	Males	6.77 (0.17)	-0.29 (0.01)	0.32 (0.01)	0.74 (0.01)	1.31 (0.02)	NS, Anchor item	
	Females							
How much did pain interfere with your enjoyment of recreational activities?	Males	6.50 (0.15)	-0.32 (0.01)	0.24 (0.01)	0.63 (0.01)	1.15 (0.02)	NS, Anchor item	
	Females							
How much did pain interfere with doing your tasks away from home (e.g., getting groceries, running errands)?	Males	7.17 (0.29)	-0.08 (0.02)	0.40 (0.02)	0.80 (0.02)	1.32 (0.03)	0.8(0.363)	14.3(0.006)
	Females	7.51 (0.24)	-0.15 (0.01)	0.35 (0.01)	0.77 (0.02)	1.25 (0.02)		
How much did pain interfere with your ability to participate in social activities?	Males	8.17 (0.23)	-0.07 (0.01)	0.38 (0.01)	0.80 (0.01)	1.27 (0.02)	NS, Anchor item	
	Females							
How much did pain interfere with work around the home?	Males	7.79 (0.20)	-0.28 (0.01)	0.28 (0.01)	0.70 (0.01)	1.25 (0.02)	NS, Anchor item	
	Females							
How much did pain interfere with the things you usually do for fun?	Males	7.94 (0.32)	-0.25 (0.02)	0.28 (0.02)	0.65 (0.02)	1.15 (0.02)	0.4(0.522)	10.5(0.033)
	Females	8.21 (0.26)	-0.22 (0.01)	0.29 (0.01)	0.71 (0.01)	1.18 (0.02)		
How much did pain interfere with your enjoyment of social activities?	Males	Item not included in the analysis						
	Females							
How often did pain keep you from socializing with others?	Males	4.35 (0.10)	-0.08 (0.01)	0.35 (0.01)	0.97 (0.02)	1.69 (0.03)	NS, Anchor item	
	Females							

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters.

* Bolded entries indicate items that evidence DIF after correction for multiple comparisons; "NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Table A7:
PROMIS pain interference nine item set: IRT item parameters and DIF statistics for the age groups; the youngest age group is the reference group

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a* DIF*	*b* DIF*
How much did pain interfere with your enjoyment of life?	Age 21 - 49	6.50 (0.30)	-0.54 (0.02)	0.07 (0.02)	0.49 (0.02)	0.99 (0.03)		
	Age 50 - 64	5.61 (0.20)	-0.60 (0.02)	0.06 (0.02)	0.52 (0.02)	1.06 (0.03)	6.1 (0.013)	2.5 (0.645)
	Age 65 - 84	4.71 (0.17)	-0.57 (0.02)	0.06 (0.02)	0.55 (0.02)	1.25 (0.04)	***23.9 (<0.001)***	9.9 (0.042)
How much did pain interfere with your ability to concentrate?	Age 21 - 49	5.23 (0.24)	-0.28 (0.02)	0.27 (0.02)	0.69 (0.03)	1.19 (0.04)		
	Age 50 - 64	4.30 (0.16)	-0.20 (0.02)	0.30 (0.02)	0.80 (0.03)	1.39 (0.04)	***10.6 (0.001)***	13.0 (0.011)
	Age 65 - 84	3.75 (0.14)	-0.09 (0.02)	0.42 (0.02)	0.96 (0.03)	1.61 (0.06)	***25.7 (<0.001)***	***51.7 (<0.001)***
How much did pain interfere with your day to day activities?	Age 21 - 49	8.12 (0.42)	-0.51 (0.02)	0.09 (0.02)	0.44 (0.02)	0.93 (0.03)		
	Age 50 - 64	7.33 (0.29)	-0.56 (0.02)	0.08 (0.02)	0.49 (0.02)	1.05 (0.03)	2.8 (0.095)	9.6 (0.049)
	Age 65 - 84	6.06 (0.23)	-0.51 (0.02)	0.09 (0.02)	0.53 (0.02)	1.15 (0.03)	***16.5 (<0.001)***	14.1 (0.007)
How much did pain interfere with your enjoyment of recreational activities?	Age 21 - 49	7.47 (0.36)	-0.55 (0.02)	0.02 (0.02)	0.34 (0.02)	0.83 (0.03)		
	Age 50 - 64	6.75 (0.26)	-0.58 (0.02)	0.00 (0.02)	0.39 (0.02)	0.88 (0.02)	3.1 (0.077)	4.9 (0.297)
	Age 65 - 84	6.09 (0.22)	-0.52 (0.02)	0.01 (0.02)	0.39 (0.02)	0.96 (0.03)	***9.0 (0.003)***	8.8 (0.066)
How much did pain interfere with doing your tasks away from home (e.g., getting groceries, running errands)?	Age 21 - 49 Age 50 - 64 Age 65 - 84	7.45 (0.20)	-0.35 (0.01)	0.13 (0.01)	0.54 (0.01)	1.02 (0.02)	DIF not significant	
How much did pain interfere with your ability to participate in social activities?	Age 21 - 49 Age 50 - 64 Age 65 - 84	8.28 (0.23)	-0.31 (0.01)	0.14 (0.01)	0.55 (0.01)	1.02 (0.02)	DIF not significant	
How much did pain interfere with work around the home?	Age 21 - 49 Age 50 - 64 Age 65 - 84	7.92 (0.21)	-0.51 (0.01)	0.04 (0.01)	0.45 (0.01)	1.00 (0.02)	NS, Anchor item	
How much did pain interfere with the things you usually do for fun?	Age 21 - 49 Age 50 - 64 Age 65 - 84	8.21 (0.22)	-0.46 (0.01)	0.05 (0.01)	0.44 (0.01)	0.92 (0.02)	NS, Anchor item	
How much did pain interfere with your enjoyment of social activities?	Age 21 - 49 Age 50 - 64 Age 65 - 84	Item not included in the analysis						
How often did pain keep you from socializing with others?	Age 21 - 49	4.17 (0.19)	-0.29 (0.03)	0.13 (0.03)	0.82 (0.03)	1.53 (0.06)		
	Age 50 - 64	4.52 (0.16)	-0.33 (0.02)	0.10 (0.02)	0.70 (0.02)	1.41 (0.04)	2.4 (0.123)	8.8 (0.066)
	Age 65 - 84	4.58 (0.17)	-0.31 (0.02)	0.11 (0.02)	0.69 (0.02)	1.39 (0.05)	3.8 (0.053)	9.9 (0.042)

* Statistical test for differences in parameters is Wald test using 1 df for the test of differences in the a parameters for the comparison groups and 2 df for the test of differences in the b parameters.

* Bolded entries indicate items that evidence DIF after correction for multiple comparisons; "NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Table A8:

PROMIS pain interference nine item set: IRT item parameters and DIF statistics for the language groups for Hispanics only (n = 1,038), participants interviewed in English are the reference group and those in Spanish are the focal group

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a* DIF*	*b* DIF*
How much did pain interfere with your enjoyment of life?	English Interview	5.31 (0.33)	-0.39 (0.05)	0.22 (0.03)	0.72 (0.03)	1.25 (0.05)	1.0 (0.307)	***15.0 (0.005)***
	Spanish Interview	4.78 (0.41)	-0.26 (0.05)	0.33 (0.04)	0.82 (0.05)	1.49 (0.08)		
How much did pain interfere with your ability to concentrate?	English Interview	4.39 (0.24)	-0.03 (0.03)	0.48 (0.03)	1.01 (0.04)	1.61 (0.06)	NS, Anchor item	
	Spanish Interview							
How much did pain interfere with your day to day activities?	English Interview	7.02 (0.48)	-0.38 (0.05)	0.28 (0.03)	0.69 (0.03)	1.26 (0.04)	1.4 (0.243)	14.0 (0.007)
	Spanish Interview	6.16 (0.55)	-0.19 (0.05)	0.32 (0.04)	0.65 (0.04)	1.23 (0.06)		
How much did pain interfere with your enjoyment of recreational activities?	English Interview	7.25 (0.48)	-0.35 (0.04)	0.15 (0.03)	0.58 (0.03)	1.10 (0.04)	0.1 (0.713)	***21.4 (<0.001)***
	Spanish Interview	6.96 (0.62)	-0.20 (0.05)	0.31 (0.04)	0.72 (0.04)	1.26 (0.06)		
How much did pain interfere with doing your tasks away from home (e.g., getting groceries, running errands)?	English Interview	6.58 (0.38)	-0.17 (0.04)	0.31 (0.03)	0.73 (0.03)	1.30 (0.04)	NS, Anchor item	
	Spanish Interview							
How much did pain interfere with your ability to participate in social activities?	English Interview	7.64 (0.47)	-0.14 (0.03)	0.34 (0.02)	0.75 (0.02)	1.27 (0.04)	NS, Anchor item	
	Spanish Interview							
How much did pain interfere with work around the home?	English Interview	7.11 (0.42)	-0.32 (0.04)	0.22 (0.03)	0.64 (0.02)	1.23 (0.04)	NS, Anchor item	
	Spanish Interview							
How much did pain interfere with the things you usually do for fun?	English Interview	7.58 (0.47)	-0.23 (0.04)	0.25 (0.03)	0.66 (0.02)	1.17 (0.03)	NS, Anchor item	
	Spanish Interview							
How much did pain interfere with your enjoyment of social activities?	English Interview	Item not included in the analysis						
	Spanish Interview							
How often did pain keep you from socializing with others?	English Interview	3.96 (0.21)	-0.16 (0.04)	0.24 (0.03)	0.94 (0.03)	1.64 (0.06)	NS, Anchor item	
	Spanish Interview							

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters.

* Bolded entries indicate items that evidence DIF after correction for multiple comparisons; "NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Figure A1:
PROMIS pain interference ten item set: Scree plot from exploratory factor analysis of the total sample (n = 5,475)

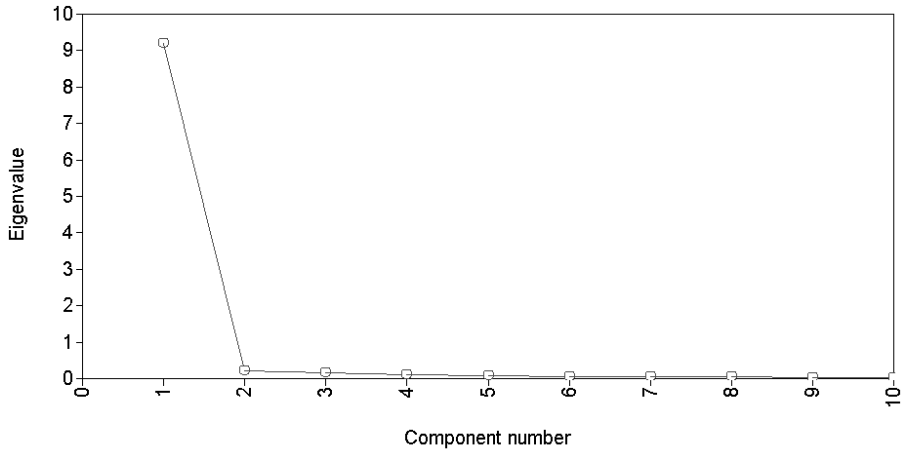


Figure A2:
PROMIS pain interference nine item set: Item information functions (Total sample)