

Establishing a common metric for self-reported pain: linking BPI Pain Interference and SF-36 Bodily Pain Subscale scores to the PROMIS Pain Interference metric

Karon F. Cook¹ · Benjamin D. Schalet¹ · Michael A. Kallen¹ · Joshua P. Rutsohn¹ · David Cella¹

Accepted: 7 April 2015 / Published online: 18 April 2015
© Springer International Publishing Switzerland 2015

Abstract

Purpose The study purposes were to mathematically link scores of the Brief Pain Inventory Pain Interference Subscale and the Short Form-36 Bodily Pain Subscale (legacy pain interference measures) to the NIH Patient-Reported Outcome Measurement Information System (PROMIS[®]) Pain Interference (PROMIS-PI) metric and evaluate results.

Methods Linking was accomplished using both equipercentile and item response theory (IRT) methods. Item parameters for legacy items were estimated on the PROMIS-PI metric to allow for pattern scoring. Crosswalk tables also were developed that associated raw scores (summed or average) on legacy measures to PROMIS-PI scores. For each linking strategy, participants' actual PROMIS-PI scores were compared to those predicted based on their legacy scores. To assess the impact of different sample sizes, we conducted random resampling with replacement across 10,000 replications with sample sizes of $n = 25, 50,$ and 75 .

Results Analyses supported the assumption that all three scales were measuring similar constructs. IRT methods produced marginally better results than equipercentile linking. Accuracy of the links was substantially affected by sample size.

Conclusions The linking tools (crosswalks and item parameter estimates) developed in this study are robust methods for estimating the PROMIS-PI scores of samples

based on legacy measures. We recommend using pattern scoring for users who have the necessary software and score crosswalks for those who do not.

Keywords Pain · Pain measurement · Patient outcome assessment · Psychometric methods/scaling · Item response theory · Instrument calibration/equivalency among scales

Introduction

The Pain Taxonomy of the International Association for the Study of Pain defines pain as an “unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage” [1]. Unrelieved pain is recognized as “a major global healthcare problem” [2]. A common symptom of many chronic conditions, pain is not only highly prevalent [3], but it exacts a substantial toll on quality of life by interfering with mental, physical, and social activities [4–7]. Because of its prevalence, impact, and relevance to patients, pain is increasingly used as a primary or secondary outcome in clinical trials [8].

A plethora of measures of pain are in use, complicating comparisons across studies. In response, researchers have used linking methodologies to create crosswalk tables that associate scores from one pain measure to the corresponding scores of another. In a sample of older adults, Edelen and Saliba [9] linked scores from a 0–10 pain intensity item to verbal descriptors of pain. Using a sample of individuals living with multiple sclerosis, Askew et al. [10] associated scores from the Brief Pain Inventory Pain Interference (BPI-PI) Subscale [11] to the metric of the Patient-Reported Outcomes Measurement Information System (PROMIS) Pain Interference (PROMIS-PI) metric.

✉ Karon F. Cook
karon.cook@northwestern.edu

¹ Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 625 Michigan Ave, 27th Floor, Chicago, IL, USA

A shortcoming of these studies is that they were completed using narrow demographic and clinical subsamples and applied a single linking methodology. The purpose of the current study was to link scores of the BPI-PI [11] and SF-36 Bodily Pain (SF36-BP) Subscale [12] to the PROMIS-PI metric. We tested the robustness of the linkings by comparing results across multiple linking strategies and recommended the best-performing methods for linking scores to the PROMIS-PI T-score metric. Additionally, we compared the results we obtained linking BPI-PI scores to the PROMIS-PI metric to results obtained by Askew and colleagues, whose data were limited to individuals living with MS [10].

This study is part of a larger body of work aimed to produce multiple PRO crosswalks that link similar instruments to a common metric, creating a Rosetta Stone linkage (PROsetta Stone[®]; 1RC4CA157236-01, PI: David Cella). The primary aim of the PROsetta Stone[®] project was to link the scores of “legacy measures” to PROMIS’ score metric. The PROMIS metric uses the T-score metric, which has a mean of 50 and a standard deviation of 10. The metric is anchored at the mean of a sample matched to the 2000 US general population census with respect to marginal distributions for gender, age, race/ethnicity, education, marital status, and income [13]. The centering of the metric to the US general population creates a convenient interpretive context for scores. For example, a person receiving a PROMIS-PI T-score of 60 can be interpreted as having worse pain, by 1 SD, compared to the US “average.” Though the general population norms are embedded within the PROMIS metric, norms for subpopulations also can be generated [14, 15].

Methods

Measures

PROMIS Pain Interference

Measures of pain interference quantify the impact of pain on a wide range of life activities including social function, physical function, work, recreation, leisure, family roles, activities of daily living, and sleep. The PROMIS Pain Interference item bank consists of 41 items with a 7-day time frame. It uses a 5-point rating scale (1—“not at all” to 5—“very much”), with higher scores indicating more pain interference. Details of the development of the bank and its psychometric evaluation have been published [16]. Briefly, a database of items was created that included published items and items written based on feedback from patients. BPI and SF36-BP items were included in the database but not included in the candidate item bank because they are

proprietary measures and because their response scales are inconsistent with those selected for PROMIS items. Candidate items were evaluated based on patient interviews and review by clinical experts in pain. Item responses were calibrated using the graded response model [17], an item response theory (IRT) model appropriate for modeling item responses that have more than two ordered response categories. Tests of model fit, differential item function, precision, and validity were conducted, and the item bank was reduced on the basis of the results. Findings supported the psychometric soundness of the item bank. Calculation of trait-level-specific test information indicated reliability greater than 0.95 for all levels of pain interference except for extremely low levels. Correlations with scores of other measures supported the concurrent validity of PROMIS-PI scores (e.g., 0.78 with BPI; 0.73 with SF-36 BP) and their discriminant validity (e.g., 0.48, 0.35, and 0.33 with scores on PROMIS measures of fatigue, anxiety, and depression, respectively) [16]. The content and response options of all PROMIS items and items of other measures are reported in “Appendix 1.”

Brief Pain Inventory Pain Interference Subscale

Substantial evidence has accumulated for both the reliability and validity of the BPI [11, 18, 19]. For example, the BPI-PI has been found to correlate highly with the WHO Disability Assessment Schedule 2.0 (0.69–0.81, from 1 to 12 months) [20], to discriminate between recurrent and non-fallers (sensitivity = 84.4 % and specificity = 57.8 % with a cutoff score of 4.6) [21], and to have good inter-item consistency ($\alpha = 0.89$) and responsiveness (standardized response mean = 0.91) [22]. The measure has been used in a substantial range of diseases and conditions and has been translated into many languages [23]. Developed using traditional methods, the BPI produces pain severity and pain interference scores that range from 0 to 10, with higher scores indicating worse pain. The interference subscale, BPI-PI, consists of seven items scored on an 11-point response scale that ranges from 0 = “no interference” to 10 = “complete interference” (see “Appendix 1”). Scores are computed as the sum of all item responses (range 0–70). The context for the items is “average pain” over the past week.

Short Form 36 (SF-36) Bodily Pain Subscale

The SF-36 is a 36-item health survey comprised of eight subscales measuring functional health and well-being [24]. By 2000, the SF-36 had been cited in more than 1000 publications [25]; its psychometric properties have been evaluated across many diseases and conditions [26]. Among the eight SF-36 subscales is the Bodily Pain

Subscale (SF-36 BP), which consists of two items: (a) “How much bodily pain have you had in the past 4 weeks?” and (b) “during the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?” The first item is scored from 1—“None” to 6—“Very severe”, and the second is scored from 1—“Not at all” to 5—“Extremely” (see “Appendix 1”). For the current study, we scored the SF36-BP as the sum of the raw item scores, resulting in a range of 2–11.

Samples

Data for use in the PROsetta linking studies came primarily from secondary data sources. For some linking projects, new data were collected. For the current study, only data collected for calibration of the first PROMIS item banks were used [16]. The PROMIS sample was recruited from the US general population by internet panel survey providers and from clinical populations by individual PROMIS investigators. In the PROMIS data collection strategy, a subsample of participants responded to every candidate item of a PROMIS measure as well as to all items of one or more “legacy instruments” that measured the same or a similar domain. Individuals who responded to the PROMIS-PI candidate items also completed the BPI-PI and the SF-36 BP.

After evaluating the candidate PROMIS-PI items, 41 were retained for the calibrated bank [16]. For the current study, we have included in each linking sample only data from persons who completed all 41 of these items and all the items of the respective linked scales (BPI-PI and SF-36 BP).

Analyses

Assumption tests

Details of the PROsetta analytic strategy have been published [27]. The approach begins by testing several linking assumptions. One of the assumptions is that scores from the measures to be linked are strongly associated. We evaluated this assumption using correlational analyses. Dorans recommended a threshold $r \geq 0.86$ between scores on two measures as indicating strong enough association for scale linking [27]. We also calculated the item-to-total correlations (adjusted for overlap) for the combined set of PROMIS and legacy items.

A second linking assumption is that scores on the scales to be linked measure the same (or very similar) constructs. This is also an assumption for employing some of the linking methods used in the study, i.e., those requiring calibration to an IRT model, specifically, the graded

response model [17]. Unidimensionality of the data was evaluated using confirmatory and exploratory factor analyses on combined item response data (PROMIS-PI and linked scale items). In confirmatory factor analyses (CFA), all items of the PROMIS-PI and the legacy instrument were modeled as loading on a single factor. These analyses were conducted using the WLSMV estimator of MPlus [28]. Because MPlus allows a maximum of ten response categories per item, BPI-PI responses (0–10) had to be recoded from eleven possible responses to ten. This was accomplished by collapsing the top two response categories (9 and 10) into a single category. Polychoric correlations were used to account for the ordinal nature of the data. We calculated fit statistics to help quantify the fit of a unidimensional model, including the Comparative Fit Index (CFI), the Tucker Lewis Index (TLI), and the root mean square error of approximation (RMSEA). A number of criteria have been proffered for classifying degrees of model fit based on these statistics. There is no general consensus regarding these criteria, and in fact, recommendations often are misstated [29]. Further, fit statistics can be influenced substantially by extraneous factors [30, 31]. For our purposes, we used the following criteria: $RMSEA < 0.08$ [32], $TLI > 0.95$, and $CFI > 0.95$ [33, 34]. We acknowledge the limitations of such standards; we applied them as informative guides for judging the relative unidimensionality of item responses and not as canonical benchmarks.

We also estimated omega hierarchical (ω_h), the proportion of total variance attributable to a general factor [35, 36], using the *psych* package [37] in *R* (version 1.2.8) [38]. This was accomplished using a bifactor model [39–41]. The bifactor model method estimates ω_h from the general factor loadings derived from an exploratory factor analysis and a Schmid–Leiman transformation. Values of 0.70 or higher for ω_h suggest that the item set is sufficiently unidimensional for most applications of unidimensional IRT models [41]. In addition, we calculated explained common variance (ECV), the ratio of the common variance explained by the general factor to the total common variance, using version 1.5.1 of the *psych* package [42] in *R*. The ECV estimates the relative strength of a general factor compared to group factors. Reise recommended ECV values ≥ 0.60 as a tentative benchmark.

In addition to unidimensionality, IRT calibrations assume local independence. Local independence is the assumption that, once the dominant factor is accounted for, there should be no significant associations among item responses [43]. Because the linked scales had similar content, there was concern that there could be local dependency between item pairs comprised of a PROMIS item and an item of one of the linked scales. We tested for this inter-scale, local dependency in two ways. We estimated a

unidimensional confirmatory factor analysis model and then flagged pairs of items with residual correlations >0.20 . In addition, we calculated the LD statistic [44]. Described earlier for dichotomous responses, the LD statistic was generalized for polytomous responses in IRTPRO [45]. LD statistics ≥ 10 may indicate problematic local dependency.

To further explore the strength of the relationship between PROMIS scores and scores on legacy measures, we evaluated invariance across subpopulations by calculating the root expected mean square differences (REMSDs) for subsamples. REMSD is estimated by subtracting standardized mean differences for two subgroups. Scores were categorized by gender and by age (<65 and ≥ 65 years). Dorans and Holland [46] suggested that, when less than 8 % of the total variance is explained by differences in subpopulations, invariance is supported.

Cross-method comparisons

Applying different analytic strategies is an effective way to test the sensitivity of results to linking method [47]. In the current study, two families of methods were used—equipercentile linking [47, 48] and IRT linking. Within each family of links, we conducted several variations. Good agreement among results of these methods was judged as indicating a robust linking relationship.

In equipercentile linking, a nonlinear linking relationship is estimated that matches scores from two linked scales based on their percentile ranks. For example, the median BPI-PI score for the sample would be associated with the PROMIS-PI median score for the same sample. By matching across percentile ranks, a best fitting function is derived. The equipercentile linking was conducted using the LEGS program [49]. To reduce the impact of random sampling error [47, 49, 50], equipercentile linking can be conducted in conjunction with a pre- or post-smoothing method. Pre-smoothing involves smoothing the observed score distributions from the two measures to be linked, prior to their linking; post-smoothing involves smoothing the equipercentile equating function that is a product of the unsmoothed observed score distributions. In the current study, we applied the LEGS cubic-spline post-smoothing algorithm with which a smoothing cubic-spline function is fit to the obtained equipercentile equating function, with the degree of smoothing to be conducted set by a smoothing parameter “*s*” [51]. Setting “*s*” to 0.0 creates a “no smoothing” condition, while “*s*” settings of 0.3 and 1.0 represent “less” and “more” smoothing, as defined by Brennan [49]. For the current study, we compared results based on three smoothing conditions: 0.0, 0.3, and 1.0. With this algorithm, linear interpolation is used to determine score equivalents for some extreme high and low

scores for which the smoothing cubic-spline function cannot be computed [51].

When scale data met IRT assumptions, we also used “fixed-parameter calibration.” The items for each legacy scale were combined with the PROMIS-PI items (PROMIS-PI + BPI-PI; PROMIS-PI + SF36-BP). These combined item pools were calibrated in single runs with PROMIS-PI item parameters fixed at their previously published values [16]. This approach produces parameter estimates for the items of the legacy scale that are on the same metric as PROMIS-PI scores. In IRT, scores can be estimated using any subset of items (the basis for computer adaptive testing). Using the item parameters obtained for the legacy instruments, we estimated PROMIS-PI scores based only on individuals’ patterns of responses to legacy instrument items, hereafter referred to as “IRT pattern scoring.” IRT pattern scoring requires users to have item-level data and software that can derive score estimates based on input item responses and item parameters.

We also constructed “crosswalks” to associate summed item scores on legacy measures to their most closely associated PROMIS T-scores, basing the association on the established IRT calibrations. This was accomplished by applying an expected a posteriori (EAP) summed scoring approach, which takes into account the fact that more than one response pattern can result in a given summed (or average) score. For example, there are many response patterns that would result in a score of “6” on the SF-36 BP including: an item score of “3” on both items; a score of “4” on one item and “2” on another; and the unlikely pattern of “5” on one item and “1” on another. Though all these patterns result in the same summed score, they would each have a different IRT-scaled score. Summed score EAP (SSEAP) weights the likelihood of the different response patterns for a given summed (or averaged) score and identifies its mean IRT-scaled score [52]. EAP summed scoring was used in the current study, and the results were tabulated into a crosswalk table. Hereafter, we refer to these results as “IRT crosswalk.”

Sample size comparisons

From each linking, we obtained predicted PROMIS-PI scores based on each of the legacy instruments. In addition, we had actual PROMIS-PI scores from all participants. The accuracy of each linking method was evaluated by estimating correlations between actual and predicted scores and calculating the means and standard deviations of differences in scores. To evaluate bias, estimate a standard error, and assess the impact of different sample sizes, we conducted random resampling with replacement across 10,000 replications with sample sizes of $n = 25, 50,$ and 75 . The mean of the difference scores (PROMIS-PI

observed score minus link-predicted PROMIS-PI score) was computed for each replication. The mean of the replication means and their standard deviations were used as estimates of bias and empirical standard error, respectively. After reviewing the results across linking methods, a recommended link was chosen, and crosswalk tables were constructed that associated summed scores on the BPI-PI and the SF36-BP with PROMIS-PI T-scores. Finally, the results from the BPI-PI crosswalk were compared to the crosswalk constructed by Askew and colleagues, which was also based on the fixed item parameter linking of the BPI-PI but used a more homogenous sample (individuals living with MS) and was based on a PROMIS-PI short form [10].

Results

Samples

The numbers of respondents in the samples used to link SF36-BP and BPI-PI scores to the PROMIS metric were 694 and 736, respectively. All participants answered the PROMIS items. Those answering the other two measures were not unique samples, and in fact, the SF36-BP sample was a subsample of the BPI-PI sample. Table 1 presents sample demographics by linking sample. As the table shows, there were more female than male respondents in the data sets (53.4–53.9 %). Respondents were predominately white (80.6–81.3 %) and most (82.8–82.9 %) had at least some college.

Assumption tests

Linking assumptions

Inspection of item content revealed substantial overlap among PROMIS-PI and pain legacy measures. The seven items of the BPI-PI ask about the impact of pain “during the past week” on general activity, mood, walking ability, normal work (includes work outside the home and housework), relations with other people, sleep, and enjoyment of life. In the PROMIS-PI item bank, there are analogous items that target each of these areas. For example, items in the PROMIS-PI bank ask about the impact of pain on “day to day activities,” walking, working, socializing, relationships, sleeping, and enjoyment of life. In addition, PROMIS-PI items ask about how much (or how often) respondents were tense, worried, or “felt depressed” because of pain. The SF36-BP subscale has an item about interference with work in and outside the home. As already noted, a similar item is included in both the PROMIS-PI and the BPI-PI. The other item of the SF36-BP is an

Table 1 Demographics by sample

Variable	Brief Pain Inventory Pain Interference (<i>n</i> = 736)	Short Form-36 Body Pain (<i>n</i> = 694)
Age	50.1 (18.46)	50.1 (18.58)
Gender		
Male	339 (46.1)	323 (46.5)
Female	397 (53.9)	371 (53.4)
Marital status		
Never married	135 (18.3)	130 (18.7)
Married	412 (55.9)	384 (55.3)
Co-resident with partner	61 (8.3)	60 (8.7)
Separated	9 (1.2)	8 (1.2)
Divorced	85 (11.6)	79 (11.4)
Widowed	34 (4.6)	33 (4.8)
Education		
Eighth grade	2 (0.27)	1 (0.14)
Some high school	15 (2.0)	13 (1.8)
High school grade/GED	110 (14.9)	105 (15.1)
Some college	241 (32.8)	226 (32.6)
College degree	214 (29.1)	202 (29.1)
Advanced degree	154 (20.9)	147 (21.2)
Income		
Less than \$20,000	69 (9.6)	61 (9.0)
Between \$20k and \$50k	253 (35.3)	239 (35.4)
Between \$50k and \$100k	279 (38.9)	270 (39.9)
\$100k or more	116 (16.2)	106 (15.7)
Race		
White	593 (80.6)	564 (81.3)
Black	62 (8.4)	54 (7.8)
Asian	3 (0.41)	3 (0.43)
Native American	3 (0.41)	3 (0.43)
Pacific Islander	2 (0.27)	1 (0.14)
Other	38 (5.2)	37 (5.3)
Multiracial	35 (4.8)	32 (4.6)

intensity item, “How much bodily pain have you had in the past 4 weeks?” Though PROMIS and the BPI measure pain intensity using different scales, pain intensity and pain interference are strongly correlated pain domains [53].

Table 2 provides item and scale score correlations for the PROMIS-PI, the two legacy scales, and the combined PROMIS and legacy scale items. The unadjusted correlations between PROMIS-PI T-scores and legacy instrument scores were above (BPI-PI: $r = 0.93$) or just slightly below (SF36-BP: $r = 0.84$) Dorans’ recommended threshold ($r \geq 0.86$). [54]. Item-to-total correlation estimates (adjusted for overlap) were high for the 41 PROMIS-PI items alone (range 0.59–0.89) and when combined with legacy items (range 0.59–0.90).

IRT assumptions

Results relevant to tests of dimensionality are included in Table 2. For the combined item sets of PROMIS and legacy items, CFA fit statistics ranged from adequate to very good, depending on the fit statistic referenced (see Table 2). The combined PROMIS-PI and BPI-PI (48 items) fit values were: CFI = 0.951, TLI = 0.948, and RMSEA = 0.093. For PROMIS and SF36-BP (43 items), fit values were: CFI = 0.97, TLI = 0.97, and RMSEA = 0.081. These results suggest essential unidimensional data-model fit. High values of ω_h estimates, 0.83 and 0.84 for BPI-PI and SF36-BP, respectively, suggest the presence of a dominant general factor for each instrument pair [55]. ECV values also were high at 0.76 and 0.78 for BPI-PI and SF36-BP items combined with PROMIS-PI items, respectively.

We also tested for local dependence between pairs of items representing PROMIS and linked scale items. This analysis was conducted with both combined items sets—the 41 PROMIS-PI items combined with the two SF36-BP items and the 41 PROMIS-PI items combined with the seven BPI-PI items. No cross-scale item pair had residual correlations exceeding 0.20, nor did any have an LD Chi-square statistic value ≥ 10.0 .

Low REMSD values confirmed invariance of scores across gender and age subgroups. For the BPI-PI and SF36-BP, gender differences accounted for 0.6 and 0.2 % of the variance, respectively. Age accounted for 3.9 and 1.6 % of the variance in BPI-PI and SF36-BP scores, respectively. These values are well below the recommended cutoff of ≤ 8 % recommended by Dorans and Holland [46].

Accuracy comparisons

Comparison of linking methods Table 3 presents the results from our comparisons between linked scores and actual PROMIS T-scores. The method labeled “IRT pattern

scoring” refers to IRT scoring based on item parameter estimates and the pattern of responses given by individuals to those items; it requires a software program and use of the item parameter estimates included in the appendices. The appendices report legacy instrument item parameters as obtained from the fixed-parameter IRT calibrations.

IRT pattern scoring and crosswalk scoring provided the most successful links between BPI-PI and PROMIS-PI scales (Table 3). The correlations between linked and actual scores were >0.90 for all methods, but the IRT pattern score resulted in the smallest mean differences in scores. However, both IRT links and all three equipercentile links also produced good results, with RMSD values around 4. The linkings between SF36-BP and PROMIS-PI were slightly less successful than those for the BPI-PI. For example, correlations between BPI-PI linked and actual PROMIS-PI scores were approximately 0.90 for all methods, compared to approximately 0.85 for SF36-BP linked scores. Comparison of the SF36-BP linking strategies also revealed differences. The IRT methods produced the most highly correlated results, the lowest RMSD values, and the least variation in difference scores (SD difference). Smoothing tended to reduce the accuracy of the equipercentile links. This may be because short form scores have few possible values, thus increasing the impact of smoothing.

The crosswalks for both legacy pain instruments are displayed in the appendices. BPI-PI summed scores (“Appendix 1”) and SF36-BP raw summed scores (“Appendix 2”), along with their corresponding PROMIS T-scores, are presented. Standard errors associated with the scaled scores also are reported.

Impact of sample size To evaluate the impact of sample size on linked score estimates, we conducted random resampling with replacement across 10,000 replications with sample sizes of $n = 25, 50, \text{ and } 75$. The findings are presented in Table 4. Recall that the mean differences reported in the tables are calculated by first computing the

Table 2 Correlational and dimensionality analysis results

	PROMIS-PI	BPI-PI	PROMIS-PI + BPI-PI	SF36-BP	PROMIS-PI + SF36-BP
Correlational analyses					
Cronbach’s alpha	0.99	0.95	0.98	0.82	0.99
Item-to-total r (corrected for overlap)	0.61–0.89	0.61–0.89	0.61–0.89	0.71–0.71	0.59–0.90
Unadjusted Pearson correlation with PROMIS-PI		0.93		0.84	
Disattenuated Pearson correlation with PROMIS-PI		0.96		0.93	
CFA					
Omega hierarchical (ω_h)			0.83		
CFI			0.97		0.97
TLI			0.97		0.97
RMSEA			0.082		0.081

Table 3 Comparison of linking methods

	Correlation	Mean difference	Standard deviation of difference	Root mean squared deviation
BPI-PI to PROMIS-PI				
IRT pattern scoring	0.904	−0.012	4.014	4.014
IRT crosswalk	0.904	0.188	4.007	4.012
EQP raw-scale SM = 0.0	0.902	0.271	4.120	4.129
EQP raw-scale SM = 0.3	0.903	0.173	4.093	4.096
EQP raw-scale SM = 1.0	0.904	0.120	4.039	4.041
SF36-BP to PROMIS-PI				
IRT pattern scoring	0.860	−0.165	4.667	4.668
IRT crosswalk	0.852	−0.132	4.785	4.787
EQP raw-scale SM = 0.0	0.850	−0.095	4.860	4.861
EQP raw-scale SM = 0.3	0.840	0.720	5.445	5.492
EQP raw-scale SM = 1.0	0.836	0.958	5.678	5.759

BPI-PI Brief Pain Inventory Pain Interference, *PROMIS-PI* Patient-Reported Outcomes Measurement Information System Pain Interference Scale, *SF36-BP* Short Form-36 Bodily Pain, *IRT* item response theory, *EAP* expected a priori, *EQP* equipercentile

Table 4 Summary of results from resampling with replacement (10,000 samples per sample size condition)

	Samples of 25		Samples of 50		Samples of 75	
	Mean differences	Standard deviation of mean differences	Mean differences	Standard deviation of mean differences	Mean differences	Standard deviation of mean differences
BPI-PI to PROMIS-PI						
IRT pattern scoring	−0.004	0.787	−0.010	0.553	−0.013	0.445
IRT crosswalk	0.184	0.792	0.190	0.539	0.188	0.438
EQP raw-scale SM = 0.0	0.264	0.815	0.278	0.570	0.269	0.453
EQP raw-scale SM = 0.3	0.173	0.789	0.180	0.563	0.179	0.450
EQP raw-scale SM = 1.0	0.135	0.795	0.126	0.547	0.119	0.441
SF36-BP to PROMIS-PI						
IRT pattern scoring	−0.149	0.923	−0.177	0.629	−0.169	0.506
IRT crosswalk	−0.144	0.931	−0.127	0.654	−0.123	0.522
EQP raw-scale SM = 0.0	−0.105	0.962	−0.099	0.657	−0.090	0.530
EQP raw-scale SM = 0.3	0.723	1.069	0.707	0.741	0.730	0.602
EQP raw-scale SM = 1.0	0.959	1.131	0.966	0.759	0.962	0.615

BPI-PI Brief Pain Inventory Pain Interference, *PROMIS-PI* Patient-Reported Outcomes Measurement Information System Pain Interference, *SF-36* Short Form-36 Bodily Pain, *IRT* item response theory, *EQP* equipercentile

mean difference within each of the 10,000 samples for a given sample size. Next, the mean of these means was calculated. The results reported in Table 4 center on the “mean differences” for each sample size condition. As expected, accuracy of linked scores was better for larger sample sizes. The bigger gain in accuracy was for a sample size increase from $n = 25$ – 50 (compared to the increase from $n = 50$ – 75). The distribution of mean differences indicated little bias (<1 T-score unit for all sample sizes, all methods, and both scales).

Of greater relevance to future use of the linking results is the variability of results by linking method obtained across the 10,000 replications with different sample sizes. For the BPI-PI link, IRT crosswalk scoring resulted in the lowest

empirical standard deviations (e.g., 0.439 for $n = 75$), followed by equipercentile link with the most smoothing (e.g., 0.439 for $n = 75$). The trend was slightly different for the SF36-BP link. For sample sizes of $n = 75$, the standard deviations were lowest for IRT pattern scoring and next lowest for IRT crosswalk scores. Note that these standard errors can be used to create confidence intervals around linking results. That is, if the PROsetta Stone crosswalk tables were used to estimate PROMIS-PI scores from BPI-PI scores, there would be a 95 % probability that the difference between the mean of this linked PROMIS-PI T-score and the mean of the actual PROMIS-PI T-score (if obtained) would be within ± 1.53 , 1.07, and 0.86 T-score units, respectively, for samples sizes of $n = 25$, 50, and 75 (i.e., $1.96 \times \text{SD}$ of

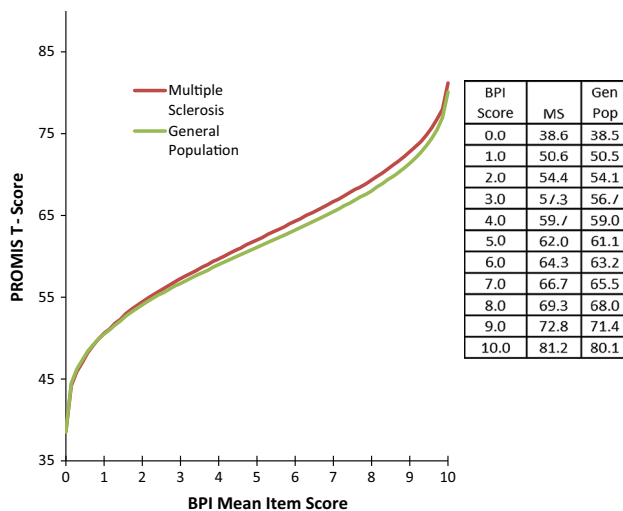


Fig. 1 Comparison of crosswalk results based on the fixed-parameter linking in a sample of individuals with multiple sclerosis (*MS*) and the general population (*Gen Pop*) sample used in the current study. Table compares crosswalk values from both studies for whole number BPI-PI scores

mean differences). [For SF36-BP, the values would be: 1.82, 1.28, and 1.02.] These ranges compare favorably with the estimated minimally clinically important difference for the PROMIS-PI of 4–6 points [56].

Comparison of BPI-PI link to MS link The results of Askew et al. [10] in linking BPI-PI scores to the PROMIS-PI metric were compared with the results from the current study. Figure 1 is a plot of the crosswalk results from both studies. The figure includes a table that compares PROMIS-PI crosswalked values for whole number BPI-PI scores from 0 through 10. Differences between crosswalk scores were small ($REMSD = 0.93$), ranging from 0.00 to 1.40.

Discussion

We obtained substantial consistency between results based on IRT and those based on equipercenile linking, suggesting robustness of the results. The similarity of the BPI-PI linking results to those obtained by Askew and colleagues suggests robustness across samples. This was particularly heartening since the previous sample was clinical (individuals with MS), and our sample was drawn from the general population. However, linking studies should be conducted in additional populations to further define the generalizability of these results.

Taken as a whole, the links obtained based on IRT pattern scoring and the IRT crosswalks were superior for both the SF36-BP and the BPI-PI. For researchers who have access to the necessary software and access to item-level data (i.e., scores for every item, not just total scores), we recommend

use of this method to link scores to the PROMIS-PI T-score metric because missing data can be handled without imputation. When such resources are not available, the IRT crosswalk is recommended. Though equipercenile links without smoothing produced good results, IRT crosswalk scoring results were better, especially when means across samples of $n = 25, 50$, and 75 were compared.

Our study had a number of strengths. We used a single-group design, which produces more robust links [57]; multiple methods were compared; calibrations were anchored on externally derived estimates [16]. Despite these strengths, scores linked to the PROMIS metric based on legacy scores will have more error than scores obtained directly from the PROMIS-PI measure since linking error is added to measurement error. This error is mitigated in larger sample sizes, but estimates based on samples of less than $n = 50$ may not be adequate for some purposes. Further, though we compared multiple linking methods, not every approach was applied. Recently, Thissen et al. [58] have proposed the use of calibrated projection, a method that accounts for item score association due to similar wordings. We anticipate future research in which this promising method is compared to the linking methods applied in the current study.

Our resampling analysis allowed us to estimate the error associated with different sample sizes, but a more precise approach is to evaluate the robustness of the linking relationship in an independent sample. We were able to compare the BPI-PI crosswalk results because of prior work but had no such comparison for the SF36-BP measure. The resampling technique may underestimate the error introduced by linking.

In conclusion, this is the first study in health measurement to link multiple legacy measures to the PROMIS-PI metric. Based on the results, we constructed tools researchers can use to link scores from BPI-PI or the SF36-BP to the PROMIS-PI metric—parameter estimates for the items of each scale calibrated to the PROMIS-PI metric and crosswalks that associate legacy scores to the PROMIS-PI metric. Future studies will use similar methods to construct and evaluate score links to the PROMIS metric. The resulting tools will substantially increase researchers' ability to compare results across studies that used different instruments to measure the same health outcome.

Acknowledgments This research was part of the PROsetta Stone[®] project, which was funded by the National Institutes of Health/National Cancer Institute grant RC4CA157236 (David Cella, PI). For more information on PROsetta Stone, see www.prosettastone.org.

Appendix 1

See Table 5.

Table 5 Items included in calibration

Variable name	Item stem	Responses	Discrimination	Category boundaries
PROMIS PAININ1	In the past 7 days, how difficult was it for you to take in new information because of pain?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	3.34	0.84, 1.40, 2.02, 2.70
PROMIS PAININ3	In the past 7 days, how much did pain interfere with your enjoyment of life?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.98	0.13, 0.88, 1.38, 1.91
PROMIS PAININ5	In the past 7 days, how much did pain interfere with your ability to participate in leisure activities?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	5.05	0.24, 0.91, 1.45, 2.04
PROMIS PAININ6	In the past 7 days, how much did pain interfere with your close personal relationships?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.06	0.63, 1.17, 1.67, 2.20
PROMIS PAININ8	In the past 7 days, how much did pain interfere with your ability to concentrate?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	3.75	0.40, 1.11, 1.69, 2.34
PROMIS PAININ9	In the past 7 days, how much did pain interfere with your day to day activities?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	6.53	0.16, 0.9, 1.44, 2.01
PROMIS PAININ10	In the past 7 days, how much did pain interfere with your enjoyment of recreational activities?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	5.15	0.13, 0.79, 1.26, 1.85
PROMIS PAININ11	In the past 7 days, how often did you feel emotionally tense because of your pain?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	3.74	0.33, 1.02, 1.49, 2.15
PROMIS PAININ12	In the past 7 days, how much did pain interfere with the things you usually do for fun?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	5.30	0.18, 0.83, 1.29, 1.88
PROMIS PAININ13	In the past 7 days, how much did pain interfere with your family life?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	5.04	0.47, 1.08, 1.63, 2.17
PROMIS PAININ14	In the past 7 days, how much did pain interfere with doing your tasks away from home (e.g., getting groceries, running errands)?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.83	0.43, 1.00, 1.46, 2.04
PROMIS PAININ16	In the past 7 days, how often did pain make you feel depressed?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	3.18	0.42, 1.01, 1.71, 2.28
PROMIS PAININ17	In the past 7 days, how much did pain interfere with your relationships with other people?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.70	0.58, 1.19, 1.71, 2.32
PROMIS PAININ18	In the past 7 days, how much did pain interfere with your ability to work (include work at home)?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.62	0.25, 0.9, 1.40, 1.88
PROMIS PAININ19	In the past 7 days, how much did pain make it difficult to fall asleep?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	2.83	0.23, 0.95, 1.45, 2.08
PROMIS PAININ20	In the past 7 days, how much did pain feel like a burden to you?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.28	0.11, 0.76, 1.19, 1.72
PROMIS PAININ22	In the past 7 days, how much did pain interfere with work around the home?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	5.40	0.17, 0.84, 1.33, 1.96
PROMIS PAININ24	In the past 7 days, how often was pain distressing to you?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	3.60	−0.01, 0.62, 1.32, 2.08
PROMIS PAININ26	In the past 7 days, how often did pain keep you from socializing with others?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	4.88	0.58, 1.09, 1.68, 2.52

Table 5 continued

Variable name	Item stem	Responses	Discrimination	Category boundaries
PROMIS PAININ29	In the past 7 days, how often was your pain so severe you could think of nothing else?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	3.43	0.64, 1.14, 1.82, 2.92
PROMIS PAININ31	In the past 7 days, how much did pain interfere with your ability to participate in social activities?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	5.91	0.47, 1.01, 1.51, 2.08
PROMIS PAININ32	In the past 7 days, how often did pain make you feel discouraged?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	3.49	0.18, 0.76, 1.47, 2.2
PROMIS PAININ34	In the past 7 days, how much did pain interfere with your household chores?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.93	0.18, 0.84, 1.36, 1.98
PROMIS PAININ35	In the past 7 days, how much did pain interfere with your ability to make trips from home that kept you gone for more than 2 h?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.14	0.75, 1.18, 1.63, 2.13
PROMIS PAININ36	In the past 7 days, how much did pain interfere with your enjoyment of social activities?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	5.70	0.34, 0.96, 1.48, 2.04
PROMIS PAININ37	In the past 7 days, how often did pain make you feel anxious?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	2.85	0.38, 1.04, 1.76, 2.55
PROMIS PAININ38	In the past 7 days, how often did you avoid social activities because it might make you hurt more?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	4.56	0.56, 0.99, 1.58, 2.32
PROMIS PAININ40	In the past 7 days, how often did pain prevent you from walking more than 1 mile?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	2.98	0.37, 0.74, 1.11, 1.55
PROMIS PAININ42	In the past 7 days, how often did pain prevent you from standing for more than 1 h?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	2.74	0.40, 0.81, 1.23, 1.74
PROMIS PAININ46	In the past 7 days, how often did pain make it difficult for you to plan social activities?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	4.59	0.45, 0.94, 1.52, 2.12
PROMIS PAININ47	In the past 7 days, how often did pain prevent you from standing for more than 30 min?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	2.75	0.34, 0.82, 1.36, 1.93
PROMIS PAININ48	In the past 7 days, how much did pain interfere with your ability to do household chores?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	4.80	0.21, 0.80, 1.34, 1.95
PROMIS PAININ49	In the past 7 days, how much did pain interfere with your ability to remember things?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	3.07	0.88, 1.42, 1.98, 2.61
PROMIS PAININ50	In the past 7 days, how often did pain prevent you from sitting for more than 30 min?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	2.83	0.73, 1.26, 1.84, 2.59
PROMIS PAININ51	In the past 7 days, how often did pain prevent you from sitting for more than 10 min?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	2.53	1.04, 1.66, 2.34, 3.11
PROMIS PAININ52	In the past 7 days, how often was it hard to plan social activities because you didn't know if you would be in pain?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	4.65	0.65, 1.07, 1.56, 2.01
PROMIS PAININ53	In the past 7 days, how often did pain restrict your social life to your home?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	4.14	0.51, 1.00, 1.58, 2.33

Table 5 continued

Variable name	Item stem	Responses	Discrimination	Category boundaries
PROMIS PAININ54	In the past 7 days, how often did pain keep you from getting into a standing position?	1 = never, 2 = once a week or less, 3 = once every few days, 4 = once a day, 5 = every few hours	2.20	1.10, 1.57, 1.97, 2.29
PROMIS PAININ55	In the past 7 days, how often did pain prevent you from sitting for more than 1 h?	1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always	2.68	0.72, 1.20, 1.75, 2.41
PROMIS PAININ56	In the past 7 days, how irritable did you feel because of pain?	1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much	3.08	0.06, 0.93, 1.59, 2.21
SF-36 Bodily Pain-Item 1	How much bodily pain have you had during the past 4 weeks?	1 = none, 2 = very mild, 3 = mild, 4 = moderate, 5 = severe, 6 = very severe	2.94	-0.92, 0.04, 0.71, 1.69, 2.69
SF-36 Bodily Pain-Item 2	During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?	1 = not at all, 2 = a little bit, 3 = moderately, 4 = quite a bit, 5 = extremely	4.31	0.08, 0.84, 1.43, 2.16
Brief Pain Inventory Item 1	Circle the one number that describes how, during the past 24 h, pain has interfered with your General Activity	0–10, where 0 = “does not interfere” and 10 = “completely interferes”	4.40	0.03, 0.34, 0.61, 0.83, 1.02, 1.21, 1.44, 1.72, 2.07, 2.40
Brief Pain Inventory Item 2	Circle the one number that describes how, during the past 24 h, pain has interfered with your Mood	0–10, where 0 = “does not interfere” and 10 = “completely interferes”	3.24	-0.09, 0.37, 0.65, 0.93, 1.07, 1.31, 1.46, 1.69, 1.98, 2.24
Brief Pain Inventory Item 3	Circle the one number that describes how, during the past 24 h, pain has interfered with your Walking Ability	0–10, where 0 = “does not interfere” and 10 = “completely interferes”	2.78	0.10, 0.38, 0.57, 0.79, 0.97, 1.14, 1.32, 1.47, 1.82, 2.10
Brief Pain Inventory Item 4	Circle the one number that describes how, during the past 24 h, pain has interfered with your Normal Work (includes both work outside the home and housework)	0–10, where 0 = “does not interfere” and 10 = “completely interferes”	4.35	0.02, 0.39, 0.6, 0.79, 0.96, 1.15, 1.34, 1.62, 1.81, 2.20
Brief Pain Inventory Item 5	Circle the one number that describes how, during the past 24 h, pain has interfered with your Relations with other people	0–10, where 0 = “does not interfere” and 10 = “completely interferes”	3.88	0.39, 0.71, 0.95, 1.11, 1.33, 1.46, 1.66, 1.89, 2.21, 2.72
Brief Pain Inventory Item 6	Circle the one number that describes how, during the past 24 h, pain has interfered with your General Activity	0–10, where 0 = “does not interfere” and 10 = “completely interferes”	2.61	-0.01, 0.39, 0.64, 0.87, 1.04, 1.16, 1.31, 1.5, 1.83, 2.18
Brief Pain Inventory Item 7	Circle the one number that describes how, during the past 24 h, pain has interfered with your Mood	0–10, where 0 = “does not interfere” and 10 = “completely interferes”	4.41	0.06, 0.41, 0.70, 0.90, 1.06, 1.21, 1.35, 1.61, 1.81, 2.12

Appendix 2

See Table 6.

Table 6 Crosswalk between Short Form-36 Bodily Pain (SF36-BP) summed scores and Patient-Reported Outcome Measurement Information System Pain Interference (PROMIS-PI) T-score metric

SF36-BP Raw score	PROMIS-P T-score	Standard error
2	37.8	6.2
3	45.1	4.5
4	49.9	4.1
5	53.3	3.9
6	56.6	3.7
7	60.1	3.6
8	63.5	3.7
9	67.0	3.7
10	71.1	4.0
11	76.0	4.6

Appendix 3

See Table 7.

Table 7 Crosswalk between Brief Pain Inventory Pain Interference (BPI-PI) summed scores and Patient-Reported Outcome Measurement Information System Pain Interference (PROMIS-PI) T-score metric

BPI-PI score	PROMIS-P IT-score	SE
0	38.5	5.7
1	44.6	3.4
2	46.3	3.2
3	47.4	3.1
4	48.4	2.9
5	49.3	2.8
6	49.9	2.8
7	50.6	2.7
8	51.1	2.8
9	51.7	2.6
10	52.2	2.6
11	52.7	2.4
12	53.2	2.3
13	53.7	2.3
14	54.1	2.2
15	54.5	2.2
16	54.9	2.1
17	55.3	2.1
18	55.7	2.1
19	56	2.1
20	56.4	2.0
21	56.7	2.0

Table 7 continued

BPI-PI score	PROMIS-P IT-score	SE
22	57.1	2.0
23	57.4	2.0
24	57.7	2.0
25	58.1	2.0
26	58.4	2.0
27	58.7	2.0
28	59	2.0
29	59.3	1.9
30	59.6	1.9
31	59.9	1.9
32	60.2	1.9
33	60.5	1.9
34	60.8	1.9
35	61.1	1.9
36	61.4	1.9
37	61.7	1.9
38	62	1.9
BPI Pain Interference score	PROMIS T-score	SE
39	62.3	1.9
40	62.6	1.9
41	62.9	1.9
42	63.2	1.9
43	63.6	1.9
44	63.9	1.9
45	64.2	1.9
46	64.5	1.9
47	64.8	1.9
48	65.2	2.0
49	65.5	2.0
50	65.8	2.0
51	66.2	2.0
52	66.6	2.0
53	66.9	2.0
54	67.3	2.0
55	67.7	2.0
56	68.1	2.0
57	68.5	2.0
58	68.9	2.1
59	69.4	2.1
60	69.9	2.1
61	70.4	2.1
62	70.9	2.1
63	71.5	2.2
64	72.1	2.2
65	72.8	2.2
66	73.5	2.3

Table 7 continued

BPI Pain Interference score	PROMIS T-score	SE
67	74.5	2.3
68	75.6	2.5
69	77.1	2.7
70	80.2	3.4

References

- IASP Task Force on Taxonomy. (1994). Part III: Pain terms—A current list with definitions and notes on usage. In H. Merskey & N. Bogduk (Eds.), *Classification of chronic pain* (pp. 209–214). Seattle, WA: IASP Press.
- Goldberg, D. S., & McGee, S. J. (2011). Pain as a global public health priority. *BMC Public Health*, *11*, 770.
- 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. *Journal of Pain*, *11*(11), 1230–1239.
- Institute of Medicine. (2012). *Institute of Medicine (US) Committee on Advancing Pain Research, Care, and Education. Relieving pain in America: A blueprint for transforming prevention, care, education, and research*. Washington, DC: National Academies Press.
- Mystakidou, K., Parpa, E., Tsilika, E., Pathiaki, M., Gennatas, K., Smyrniotis, V., et al. (2007). The relationship of subjective sleep quality, pain, and quality of life in advanced cancer patients. *Sleep*, *30*(6), 737–742.
- Ramstad, K., Jahnsen, R., Skjeldal, O. H., & Diseth, T. H. (2012). Parent-reported participation in children with cerebral palsy: The contribution of recurrent musculoskeletal pain and child mental health problems. *Developmental Medicine and Child Neurology*, *54*(9), 829–835.
- Schirbel, A., Reichert, A., Roll, S., Baumgart, D. C., Buning, C., Wittig, B., et al. (2010). Impact of pain on health-related quality of life in patients with inflammatory bowel disease. *World Journal of Gastroenterology*, *16*(25), 3168–3177.
- Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., et al. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain*, *113*(1–2), 9–19.
- Edelen, M. O., & Saliba, D. (2010). Correspondence of verbal descriptor and numeric rating scales for pain intensity: An item response theory calibration. *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *65*(7), 778–785.
- Askew, R. L., Kim, J., Chung, H., Cook, K. F., Johnson, K. L., & Amtmann, D. (2013). Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. *Quality of Life Research*, *22*(10), 2769–2776.
- Cleeland, C. S., Gonin, R., Hatfield, A. K., Edmonson, J. H., Blum, R. H., Stewart, J. A., et al. (1994). Pain and its treatment in outpatients with metastatic cancer. *New England Journal of Medicine*, *330*(9), 592–596.
- Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483.
- Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., et al. (2010). Representativeness of the patient-reported outcomes measurement information system internet panel. *Journal of Clinical Epidemiology*, *63*(11), 1169–1178.
- Cook, K. F., Molton, I. R., & Jensen, M. P. (2011). Fatigue and aging with a disability. *Archives of Physical Medicine and Rehabilitation*, *92*(7), 1126–1133.
- Molton, I., Cook, K. F., Smith, A. E., Amtmann, D., Chen, W. H., & Jensen, M. P. (2014). Prevalence and impact of pain in adults aging with a physical disability: Comparison to a US general population sample. *Clinical Journal of Pain*, *30*(4), 307–315.
- Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W. H., Choi, S., Revicki, D., et al. (2010). Development of a PROMIS item bank to measure pain interference. *Pain*, *150*(1), 173–182.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from <http://www.psychometrika.org/journal/online/MN17.pdf>.
- Cleeland, C. S., Nakamura, Y., Mendoza, T. R., Edwards, K. R., Douglas, J., & Serlin, R. C. (1996). Dimensions of the impact of cancer pain in a four country sample: New information from multidimensional scaling. *Pain*, *67*(2–3), 267–273.
- Mendoza, T. R., Chen, C., Brugger, A., Hubbard, R., Snabes, M., Palmer, S. N., et al. (2004). Lessons learned from a multiple-dose post-operative analgesic trial. *Pain*, *109*(1–2), 103–109.
- Shulman, M. A., Myles, P. S., Chan, M. T., McIlroy, D. R., Wallace, S., & Ponsford, J. (2015). Measurement of disability-free survival after surgery. *Anesthesiology*, *122*(3), 524–536.
- Stubbs, B., Eggermont, L., Patchay, S., & Schofield, P. (2014). Older adults with chronic musculoskeletal pain are at increased risk of recurrent falls and the brief pain inventory could help identify those most at risk. *Geriatrics & Gerontology International*. doi:10.1111/ggi.12357.
- Kroenke, K., Theobald, D., Wu, J., Tu, W., & Krebs, E. E. (2012). Comparative responsiveness of pain measures in cancer patients. *Journal of Pain*, *13*(8), 764–772.
- Cleeland, C. S. (2009). *The brief pain inventory user guide*. Retrieved 4/16/2015, from http://www.mdanderson.org/education-and-research/departments-programs-and-labs/departments-and-divisions/symptom-research/symptom-assessment-tools/BPI_UserGuide.pdf
- Ware, J. E., Kosinski, M., & Keller, S. D. (1994). *SF-36 physical and mental health summary scales: A users' manual*. Boston, MA: The Health Institute.
- Ware, J. E., Jr. (2000). SF-36 health survey update. *Spine*, *25*(24), 3130–3139.
- Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 health survey: Manual and interpretation guide*. Boston, MA: The Health Institute, New England Medical Center.
- Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment*, *26*(2), 513–527.
- Muthén, L. K., & Muthén, B. O. (2006). *Mplus*. Los Angeles: Muthén & Muthén.
- Lance, C., Butts, M., & Michels, L. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202–220.
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford 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*(4), 447–460.
- Browne, M. W., Cudeck, R., Bollen, K. A., & Long, K. S. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

33. Hu, L., & Bentler, P. M. (1998). Fit Indices in covariance structure modeling: Sensitivity to underparameterization model misspecification. *Psychological Methods*, 3, 424–453.
34. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55.
35. McDonald, R. P. (1999). *Test theory: A unified treatment*. New York: Psychology Press.
36. Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70, 123–133.
37. Revelle, W. (2013). *psych: Procedures for personality and psychological research (R package version 1.2.8) (computer software)*. Evanston, IL: Northwestern University. <http://cran.r-project.org/web/packages/psych/index.html>
38. R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna: Austria R Foundation for Statistical Computing. <http://www.r-project.org/>
39. Deng, N., Guyer, R., & Ware, J. E., Jr. (2015). Energy, fatigue, or both? A bifactor modeling approach to the conceptualization and measurement of vitality. *Quality of Life Research*, 24(1), 81–93.
40. Paap, M. C., Brouwer, D., Glas, C. A., Monninkhof, E. M., Forstreuter, B., Pieterse, M. E., et al. (2015). The St George's Respiratory Questionnaire revisited: A psychometric evaluation. *Quality of Life Research*, 24(1), 67–79.
41. Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73(1), 5–26.
42. Revelle, W. (2015). *psych: Procedures for personality and psychological research (version 1.5.1)*. Evanston, IL: Northwestern University.
43. Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measure*, 15, 22–29.
44. Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289.
45. Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO 2.1 for Windows*. Lincolnwood, IL: Scientific Software International Inc.
46. Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281–306.
47. Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.
48. Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics*, 7(3), 165–174.
49. Brennan, R. (2004). *Linking with Equivalent Group or Single Group Design (LEGS) (version 2.0)*. Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment (CASMA).
50. Albano, T. (2011). *Equate: Statistical methods for test score equating (R package version 1.1-4)*. <http://cran.open-source.org/web/packages/equate/equate.pdf>
51. Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik*, 10(3), 177–183.
52. Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows user's guide*. Lincolnwood, IL: Scientific Software International.
53. Fayers, P. M., Hjerstad, M. J., Klepstad, P., Loge, J. H., Caraceni, A., Hanks, G. W., et al. (2011). The dimensionality of pain: Palliative care and chronic pain patients differ in their reports of pain intensity and pain interference. *Pain*, 152(7), 1608–1620.
54. Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4), 227–246.
55. Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73(1), 5–26.
56. Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology*, 64(5), 507–516.
57. Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16(Suppl 1), 85–94.
58. Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & Dewalt, D. A. (2011). Using the PedsQL 3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). *Quality of Life Research*, 20(9), 1497–1505.