Co-calibration of self-reported measures of Pain Interference: the Patient Reported Outcomes Measurement Information System and the Brief Pain Inventory

Robert L. Askew

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2012

Committee:

Jane M. Simoni

Dagmar Amtmann

Program Authorized to Offer Degree:

Psychology

University of Washington

Abstract

Co-calibration of self-reported measures of Pain Interference: the Patient Reported Outcomes Measurement Information System and the Brief Pain Inventory

Robert L. Askew

Chair of the Supervisory Committee:
Professor Jane M. Simoni
Department of Psychology

Introduction: To help researchers in MS take advantage of the measurement properties of the PROMIS Pain Interference instruments while maintaining continuity with previous research, a co-calibration table was developed that transforms scores from the Brief Pain Inventory (BPI-PI) to scores from the PROMIS-PI short form (SF).

Methods: Data for the development of the score co-calibration were collected as part of a longitudinal study of pain and fatigue in adults (n=369) with Multiple Sclerosis (MS). Data for the co-calibration validation were collected as part of a separate longitudinal study of Aging in individuals with physical disabilities (n=360). The development of the co-calibration was a multistage procedure that included dimensionality assessment, fitting an item response theory based two-parameter logistic model, estimating difficulty and discrimination parameters for items of the BPI-PI, deriving trait scores of pain interference, and matching instrument level summary scores.

Results: The co-calibration table functioned well with a mean difference between the actual SF scores and SF scores predicted by the table of 0.51 (sd=3.9). The largest differences were observed at the low-end of pain interference where variance in measurement is typically higher and where high precision is not necessarily as important as it is at higher levels of pain interference. The co-calibration performed nearly as well in terms of predictive accuracy in the

validation dataset. With respect to residual errors from the equating procedure, approximately 80% of the predicted scores in the developmental sample were within 4 points difference (<1/2 sd) of their respective PROMIS-PI SF scores, and 70% were within 4 points in the validation sample.

Conclusions: Two measures of pain interference were co-calibrated and a reference table of corresponding scores is presented. MS researchers and clinicians can use this table to facilitate comparisons between studies and to maintain continuity with previous research.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The contents of this manuscript were developed under grants from the Department of Education, National Institute on Disability and Rehabilitation Research (NIDRR) grant numbers H133B080024 and H133B031129. However, these contents do not necessarily represent the policy of the Department of Education, and one should not assume endorsement by the federal government. This work was also funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 5U01AR052171-03 to University of Washington. Information on the "Dynamic Assessment of Patient-Reported Chronic Disease Outcomes" can be found at http://nihroadmap.nih.gov/clinicalresearch/index.asp

DEDICATION

To Elizabeth and Maggie

Introduction

A large number of self-reported questionnaires are available to assess health-related quality of life (HRQOL) and disease symptomology. The proliferation of patient reported outcome measures indicates increasing awareness of the importance of the patient perspective in clinical assessment and HRQOL in clinical research. An unintended consequence of this proliferation is that results from different studies are often not comparable, because they use different instruments to assess similar or even identical traits. In 2004, the National Institute of Health sponsored the Patient Reported Outcomes Measurement Information System (PROMIS) initiative to develop new measures of psychological, social, and functional dimensions of health states that would maximize comparability across multiple diseases and conditions while maintaining sufficient sensitivity across the entire range of the construct of interest. (1, 2) To enhance generalizability and interpretability, all PROMIS measures have been centered on the U.S general population, i.e. the mean score of 50 on all PROMIS measures represents average ability or difficulty reported by the general US population, and the standard deviation is scaled to 10. Developed using Item Response Theory (IRT), these instruments offer important advantages over previously developed instruments, as they provide more detailed item-level response characteristics, allow inspection of measurement precision and bias across all levels of the domain of interest, and facilitate administration through computerized adaptive testing (CAT). A PROMIS item bank for measuring Pain Interference (PROMIS-PI) and the short form (SF) derived from the bank are now available. To help researchers in Multiple Sclerosis take advantage of the measurement properties of the PROMIS-PI while maintaining continuity with previous research, an item response theory based co-calibration was developed and tested to transform sum scores from a the Brief Pain Inventory pain interference subscale (BPI-PI) to the PROMIS-PI short form. (3, 4)

Methods

Participants

Data for the development of the crosswalk were collected as part of a larger longitudinal study of pain and fatigue in adults with Multiple Sclerosis (MS). At the fifth assessment time point (16 months) study participants were asked to indicate whether they had bothersome recurrent pain within the previous 3 months. 81.6% (n=376) endorsed this item and responded to both the BPI-PI and the PROMIS-PI SF. Scoring was not possible for 1.5% (n=7) of respondents who did not respond to a sufficient number of items. The co-calibration table was developed using the data from the remaining 369 respondents. Data for the validation of the co-calibration were collected by a different longitudinal study of Aging with physical disabilities. Detailed descriptions of recruitment procedures are available.(5) In the third year of follow-up, 360 participants with MS completed both the BPI-PI and the PROMIS-PI SF.

Measures

The PROMIS-PI was developed using item response theory, and as a result, it can be administered in a variety of formats including computer adaptive testing, targeted fixed length instruments with item selection guided by relevance to study population, and PROMIS fixed length short forms. A total of 41 items comprise the PROMIS-PI item bank, and the bank has sound psychometric properties with respect to construct validity, fit, precision, and population invariance. (3, 6) Study participants responded to the 6-item PROMIS-PI SF, version 1. The PROMIS-PI SF is centered on the US general population average of 50 with a standard deviation of 10.

The Pain Interference subscale of the Brief Pain Inventory (BPI-PI) has demonstrated validity and reliability.(7-9) The BPI-PI was developed using traditional factor analytic methods of classical test theory and has been used in studies of pain associated with multiple chronic disease conditions.(4) Seven items comprise the BPI-PI, each with a 0 (no interference) to 10

(complete interference) response scale. The summary score represents the average of all item responses.

Statistical Analysis

Pre-calibration assessment

Dorans and Holland outlined terminology, criteria, and methods used to link scores across different measures.(10, 11) The criteria for equating are the most strict and are rarely met in practice. Nevertheless, calibration (or concordance) can still be carried out when several critical criteria have been satisfied. We evaluated criteria relating to construct equivalence, reliability, symmetry, and population invariance. More specifically, we evaluated the equal construct requirement through item-inspection and dimensionality assessment. Findings of multidimensionality indicate that the equal construct requirement is not satisfied and that application of a unidimensional IRT model would be inappropriate. Reliability was evaluated through comparisons of PROMIS-PI SF and BPI-PI internal consistency statistics. Symmetry was not of concern, because we aimed to develop a table of matched scores as opposed to matching with a linear function that is particularly problematic for the symmetry requirement. Each of these criteria can be evaluated before calibration, but population invariance must be assessed afterwards. Accordingly, population invariance is evaluated with other performance indicators as part of the calibration validation.

Equating procedure

Unidimensionality is an important assumption underlying IRT-based trait estimation, and unidimensionality of the PROMIS-PI SF and the BPI-PI was evaluated in the context of a one-factor confirmatory factor analysis. Adequacy of model fit was assessed by multiple indices including X^2, Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI). CFI and TLI values above 0.95 have served as indicators of good fit to a unidimensional construct, (12). If the

results of the one factor CFA indicated the data were not sufficiently unidimensional, a bi-factor analysis was planned to assess the impact of modeling multidimensional data with a unidimensional model.

Next, a two-parameter logistic graded response model was fit to derive item-difficulty (**β**) and discrimination **(α)** parameters for each of the BPI-PI items. The parameters for the PROMIS-PI SF items were constrained to maintain direct comparability with the US general population. The PROMIS-PI SF item parameters that center these measures have been previously established as part of the larger NIH-PROMIS initiative, and Amtmann and colleagues detail this process for the PROMIS-PI item bank. (3) Once the BPI-PI item parameters were derived, the PROMIS-PI items were removed from the model, and equivalencies in BPI summary scores and BPI-PI trait scores on the PROMIS metric were derived using expected a priori sum score estimation.(13)

The requirement of population invariance is considered one of the most important, as the utility of scale equating is directly tied to how well the equating procedure functions in varying subgroups. (11) To assess variability in the performance of this co-calibration, the standardized Root Mean Square Difference (RMSD)(11) was compared across multiple subgroups defined by gender, race, age, education level, type of MS, and self-reported mobility status. In addition, multiple F-tests were carried out to assess variability in the performance of the equating procedure by subgroups with the criterion defined as the difference between PROMIS-PI SF scores and the scores predicted from the BPI-PI. Next, precision of the co-calibration was assessed by comparing differences between the SF scores and those predicted by the BPI-PI with published Minimal Important Differences (MIDs).(14) Bland-Altman plots facilitated inspection of differences across all levels of the trait. As a last step, the co-calibration was cross-validated in an independent sample of individuals with MS. Differences between the actual SF scores and the scores predicted from BPI-PI were again examined across the continuum through inspection of scatterplots and Bland Altman plots. Data coding and preparation was performed using STATA (IC 12.1, StatCorp, College Station, TX), dimensionality assessment with Mplus (v6.11, Muthén

& Muthén, Los Angeles, CA), and all IRT-based analyses were carried out with IRTPRO v2.1 (Scientific Software International, Lincolnwood, IL).

Results

Consistent with the distribution of MS in the population, the sample of individuals with MS (n=369) from the developmental dataset was predominantly Caucasian (92%) and female (84%) with a mean age of 53 (Table 1). Over half of participants reported relapsing/remitting type of MS, nearly one-quarter reported secondary progressive MS, and average time since MS diagnosis was 12 years (range: 2 - 61). The characteristics of the validation sample were quite similar, with an exception of median age and duration of disease (4 years later for both in the validation dataset) due to the nature of the study (aging with a physical disability).

Study investigators examined item content of both instruments and concluded that pain interference has been operationalized in similar fashion across both instruments. Both instruments have items related to pain interference with respect to work tasks, socializing, day-to-day activities, and general life enjoyment. Items unique to the SF relate to concentration and recreational activities, and those unique to the BPI relate to mobility, mood, and sleep. While both instruments have unique items, none of the items were considered extrinsic to the pain interference construct. The X^2 test of model fit for the confirmatory factor analysis was statistically significant, but unidimensionality was supported by a CFI of 0.96 and TLI of 0.95. The estimates of internal consistency also supported scale calibration with nearly identical Chronbach α coefficients (PROMIS-PI SF = 0.94; BPI-PI = 0.93). Findings related to common item content, unidimensionality, and reliability all indicate that co-calibration is appropriate.

The BPI-PI and PROMIS-PI SF co-calibration is presented in Table 2. Matched scores are also presented in Figure 1a with x's representing SF scores and o's representing scores predicted from the BPI scores by using the calibration table. The shape and location of the prediction line suggest minimal if any bias related to level of pain interference. Mean differences between the PROMIS-PI SF and predicted scores are also presented as a Bland-Altman plot (Figure 2a). The overall mean difference was 0.51 (95%CI=0.11 to 0.91), and the standard deviation was 3.92. The largest differences were observed at the extreme low-end of pain

interference where variance in measurement is typically higher and where high precision is not necessarily as important as it is at higher levels of pain interference. In the validation dataset, the co-calibration performed nearly as well in terms of predictive accuracy. As in the developmental dataset, the shape and location of the prediction line provide further support that the co-calibration is unbiased with respect to location on the latent trait. (Figure 1b) The average difference was -1.47 (95%CI= -1.91 to -1.04) with a standard deviation of 4.17. These differences are also illustrated as a Bland-Altman plot in Figure 2b. With respect to error in prediction from the co-calibration, nearly 80% of the predicted scores in the developmental sample were within 4 points difference of their respective PROMIS-PI SF scores, and 90% were within 6 points. Likewise for the validation dataset, 70% were within 4 points, and 87% were within 6 points difference of their PROMIS-PI SF scores.

Subgroup comparisons indicated that RMSD estimates ranged from 0.01 to 0.06 and were well below suggested upper-bounds,(11) indicating that the table of co-calibrated scores functioned well across subgroups. In addition, residual comparisons indicated non-significant differences in predictive accuracy across all subgroups (gender: $F=0.36$, $p=0.55$; race: $F=0.73$, $p=0.39$; age: $F=0.40$, $p=0.67$; education: $F=0.48$, $p=0.82$; MS type: $F=1.22$, $p=0.30$; mobility: $F=1.38$, $p=0.25$). Low RMSD estimates and non-significant F-tests indicate that the remaining requirement of population invariance for the calibration was satisfied.

Discussion

In this study, a table was developed of co-calibrated scores for 2 instruments that measure pain interference in MS. Co-calibration within the latent variable framework offers several advantages over other methods, particularly with respect to comparability. A simpler scale-level regression-based procedure would not partition measurement error appropriately, lead to larger error in prediction, and generally do not meet the outlined requirements for score calibration (e.g. symmetry, population invariance). Additionally, estimating difficulty levels simultaneously enables meaningful comparisons of items from the BPI-PI with those from the SF, the PROMIS-PI item bank, and even the trait levels of individuals, if desired.

The Bland-Altman plots suggested that the residual differences at very low levels of pain interference (theta: 40-50) were not sufficiently accurate, so the calibration table should not be used for samples of individuals with negligible levels of pain interference. As expected, when the crosswalk was applied to a different dataset, the average difference in prediction error was greater than in the developmental dataset. The average difference between actual SF scores and the scores predicted from the BPI PI using the table was nearly three times (|-1.47| vs |0.51|) the size of the error in the developmental dataset. Nevertheless, the magnitude of these residuals in the validation dataset was not overly concerning when taking into account that the PROMIS-PI has a standard deviation of 10, and the average residual was approximately 1.5 points. Minimal clinically important differences for the PROMIS-PI have been reported to be between 4-6 points.(14) Although these were derived within the context of cancer research, and it is unclear whether these MID estimates would be similar in MS; nevertheless, these are the only MIDs published to date. Absolute differences in SF scores and predicted SF scores for a majority of the individuals in this study were within this range, even when using the more conservative lower limit. It is important to note that co-calibration tables, such as this are only recommended for group level analysis and are not intended for estimating individual scores for use in the clinical care. Nevertheless, this score co-calibration table provides researchers who are interested in

capitalizing on the benefits of PROMIS a way to incorporate the PROMIS-PI SF into their study while maintaining research continuity and the ability to examine longitudinal trajectories of pain interference.

Limitations

In this study the co-calibration was not carried out using the full PROMIS-PI item bank, so the accuracy of this table for scores obtained through computer adaptive testing or customized short forms would need to be further investigated. In addition, this co-calibration was developed specifically for matching scores of individuals with MS. While the calibration table is expected to perform well beyond this diagnostic group, its functioning in different diagnostic groups will need to be validated before broader adoption and application.

Table 1. Demographic and clinical profiles of development and validation samples

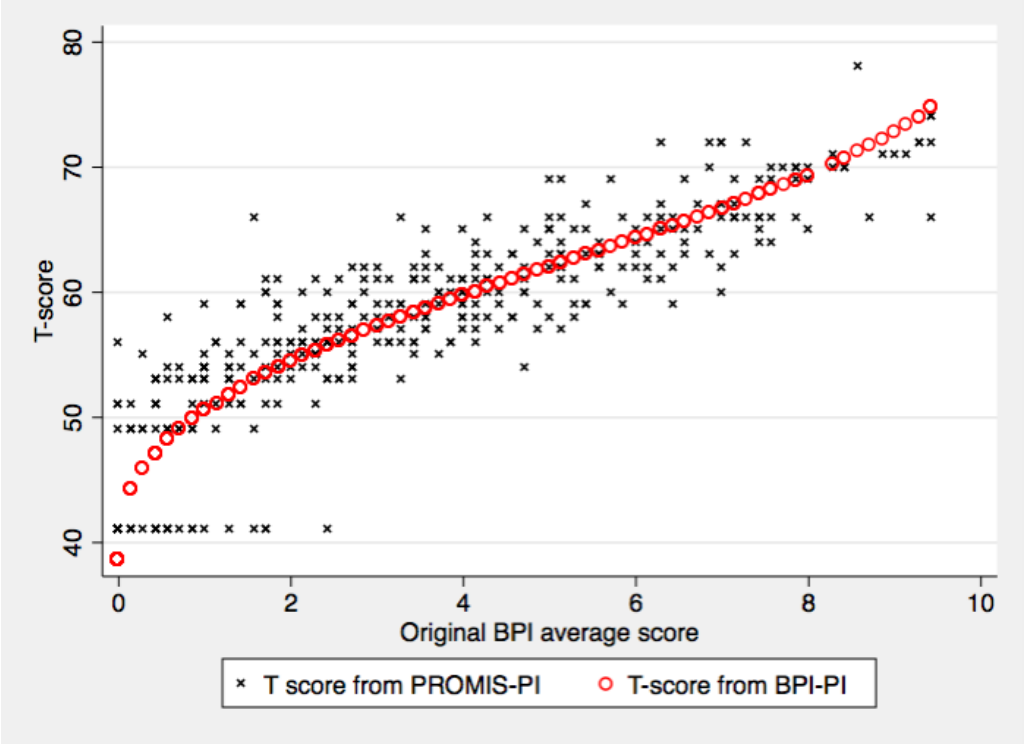
	Development dataset n=369	%	Validation dataset n=360	%
Age: mean [sd]	53.16 [10.5]		57.26 [5.5]	
Gender				
Male	60	16.3	58	16.1
Female	309	83.7	302	83.9
Ethnicity				
Caucasian	338	91.6	337	93.9
Other	31	8.4	22	6.1
Education level				
≤ High School Diploma / GED	52	14.1	35	9.8
Some college / vocational training	148	40.1	122	34.1
College graduate	106	28.7	115	32.1
Graduate or professional school	63	17.1	86	24.0
Annual household income				
< $25,000	61	16.5	54	15.0
$25,000 - 40,000	54	14.6	60	16.7
$41,000 - 55,000	45	12.2	29	8.1
$56,000 - 70,000	51	13.8	38	10.6
$71,000 - 85,000	32	8.7	25	6.9
$86,000 - 100,000	40	10.9	35	9.7
>$100,000	58	15.7	57	15.8
Decline to answer/Missing	28	7.6	62	17.2
Disease duration in years: median [range]	12 [2-61]		16 [2-41]	
Self-reported disease course				
Relapsing / remitting	203	55.0	200	55.6
Secondary progressive	86	23.3	93	25.8
Primary progressive	41	11.1	34	9.4
Progressive relapsing	34	9.2	7	2.0
Missing	5	1.4	26	7.2
Expanded Disability Status Scale (EDSS) - Mobility				
≤ 4.0	111	30.1	100	28.4
4.5 - 7.0	195	52.9	182	51.7
> 7.0	63	17.0	70	19.9

Table 2. BPI scores on PROMIS metric

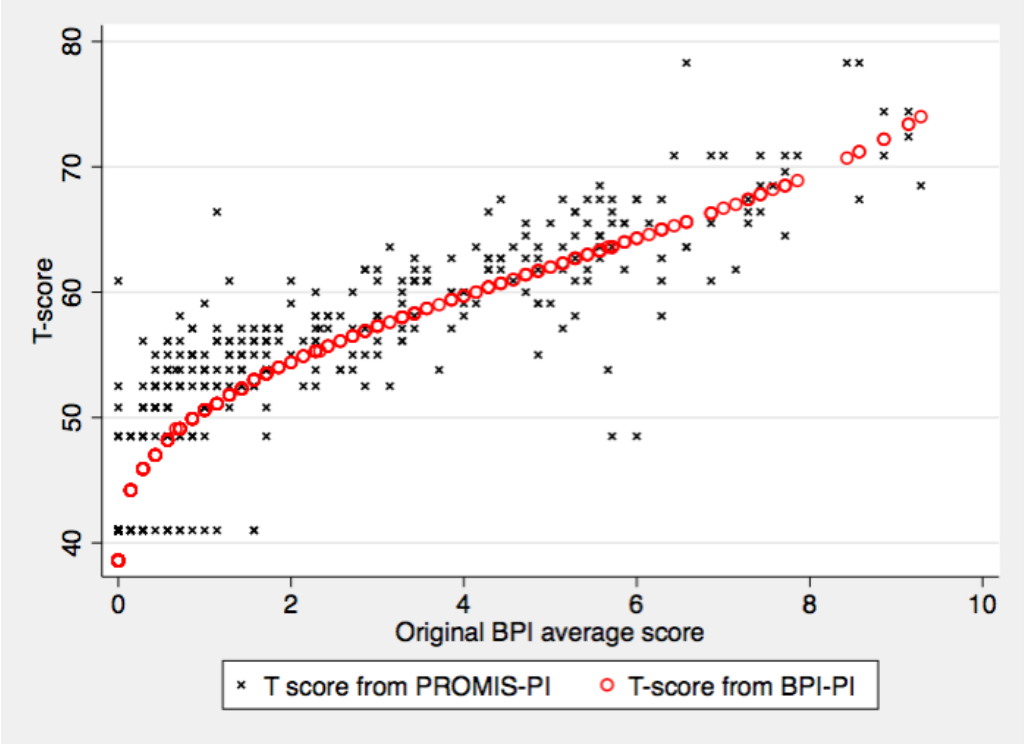
BPI Sum Score	Average BPI Score	T-Score	SD	BPI Sum Score	Average BPI Score	T-Score	SD
0	**0.0**	**38.6**	5.8	36	**5.1**	**62.3**	1.8
1	**0.1**	**44.2**	3.8	37	**5.3**	**62.7**	1.8
2	**0.3**	**45.9**	3.5	38	**5.4**	**63.0**	1.8
3	**0.4**	**47.0**	3.4	39	**5.6**	**63.3**	1.8
4	**0.6**	**48.2**	3.1	40	**5.7**	**63.6**	1.8
5	**0.7**	**49.1**	2.9	41	**5.9**	**64.0**	1.8
6	**0.9**	**49.9**	2.9	42	**6.0**	**64.3**	1.8
7	**1.0**	**50.6**	2.8	43	**6.1**	**64.6**	1.8
8	**1.1**	**51.1**	2.8	44	**6.3**	**65.0**	1.8
9	**1.3**	**51.8**	2.5	45	**6.4**	**65.3**	1.8
10	**1.4**	**52.3**	2.6	46	**6.6**	**65.6**	1.9
11	**1.6**	**53.0**	2.3	47	**6.7**	**66.0**	1.9
12	**1.7**	**53.5**	2.2	48	**6.9**	**66.3**	1.9
13	**1.9**	**54.0**	2.1	49	**7.0**	**66.7**	1.9
14	**2.0**	**54.4**	2.1	50	**7.1**	**67.0**	1.9
15	**2.1**	**54.9**	2.0	51	**7.3**	**67.4**	1.9
16	**2.3**	**55.3**	2.0	52	**7.4**	**67.8**	1.9
17	**2.4**	**55.7**	2.0	53	**7.6**	**68.2**	1.9
18	**2.6**	**56.1**	2.0	54	**7.7**	**68.5**	1.9
19	**2.7**	**56.5**	2.0	55	**7.9**	**68.9**	1.9
20	**2.9**	**56.9**	1.9	56	**8.0**	**69.3**	1.9
21	**3.0**	**57.3**	1.9	57	**8.1**	**69.8**	2.0
22	**3.1**	**57.6**	1.9	58	**8.3**	**70.2**	2.0
23	**3.3**	**58.0**	1.9	59	**8.4**	**70.7**	2.0
24	**3.4**	**58.3**	1.9	60	**8.6**	**71.2**	2.1
25	**3.6**	**58.7**	1.9	61	**8.7**	**71.7**	2.1
26	**3.7**	**59.0**	1.9	62	**8.9**	**72.2**	2.1
27	**3.9**	**59.4**	1.9	63	**9.0**	**72.8**	2.2
28	**4.0**	**59.7**	1.9	64	**9.1**	**73.4**	2.2
29	**4.1**	**60.0**	1.9	65	**9.3**	**74.0**	2.3
30	**4.3**	**60.4**	1.9	66	**9.4**	**74.8**	2.3
31	**4.4**	**60.7**	1.9	67	**9.6**	**75.7**	2.4
32	**4.6**	**61.0**	1.9	68	**9.7**	**76.8**	2.6
33	**4.7**	**61.4**	1.9	69	**9.9**	**78.0**	2.7
34	**4.9**	**61.7**	1.9	70	**10.0**	**81.2**	3.7
35	**5.0**	**62.0**	1.9				

Figures 1a/b. Scatterplots of pain interference T-scores from the PROMIS-PI Short Form (black x) and the BPI-derived estimates (red o)

1a. Developmental dataset

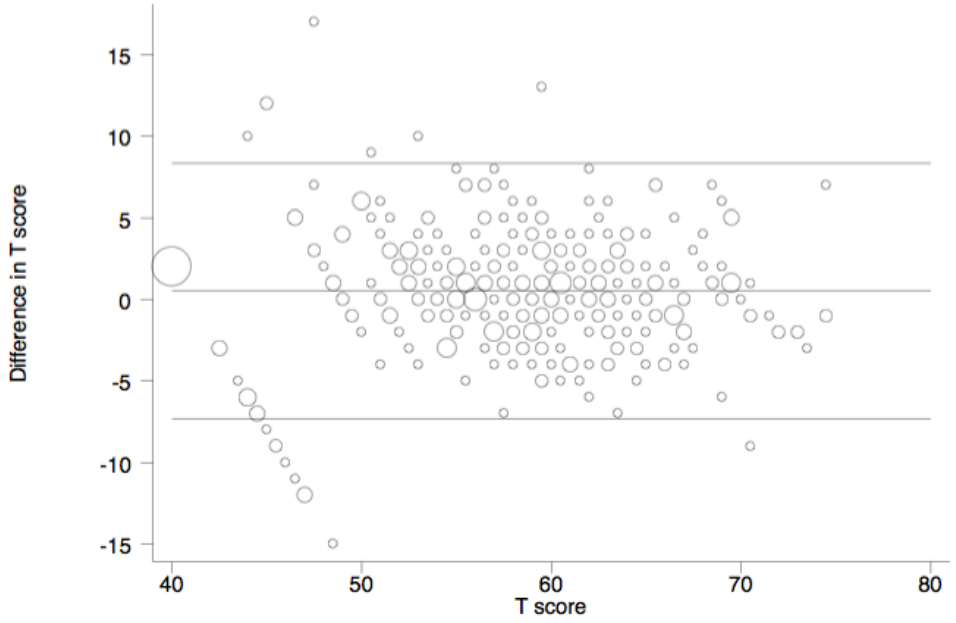


1b. Validation dataset

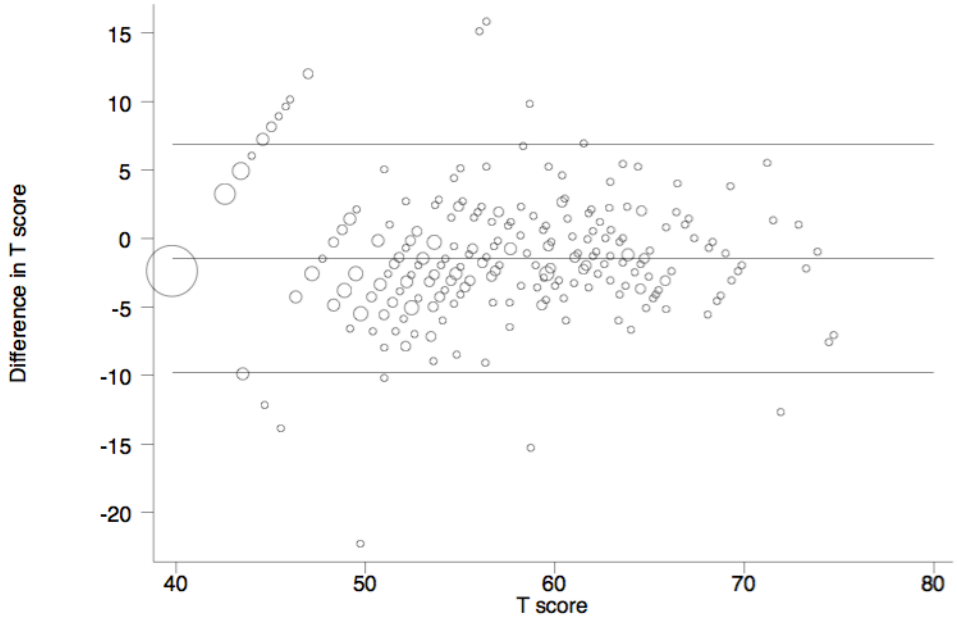


Figures 2a/b. Differences in pain interference T-scores from the PROMIS-PI Short Form and the BPI-PI derived estimates.

2a. Developmental dataset



2b. Validation dataset



REFERENCES

1. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010; 63(11):1179-94.
2. Weinfurt KP, Cella D, Forrest CB, Shulman L. NIH PROMIS 2011: Advancing PRO Science in Clinical Research and Patient Care. NIH, Bethesda, Maryland. 2011. 02:35:19.
3. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010; 150(1):173-82.
4. Cleeland CS. The Brief Pain Inventory User Guide. Houston, TX. 2009. Available from: http://www.mdanderson.org/education-and-research/departments-programs-and-labs/departments-and-divisions/symptom-research/symptom-assessment-tools/BPI_UserGuide.pdf.
5. Alschuler KN, Gibbons LE, Rosenberg DE, Ehde DM, Verrall AM, Bamer AM, et al. Body mass index and waist circumference in persons aging with muscular dystrophy, multiple sclerosis, post-polio syndrome, and spinal cord injury. Disabil Health J. 2012; 5(3):177-84.
6. Kim J, Chung H, Amtmann D, Revicki DA, Cook KF. Measurement invariance of the PROMIS pain interference item bank across community and clinical samples. Quality of life research. 2012. Epub ahead of print.
7. Mendoza TR, Chen C, Brugger A, Hubbard R, Snabes M, Palmer SN, et al. Lessons learned from a multiple-dose post-operative analgesic trial. Pain. 2004; 109(1-2):103-9.
8. Cleeland CS, Gonin R, Hatfield AK, Edmonson JH, Blum RH, Stewart JA, et al. Pain and its treatment in outpatients with metastatic cancer. The New England journal of medicine. 1994; 330(9):592-6.
9. Cleeland CS, Nakamura Y, Mendoza TR, Edwards KR, Douglas J, Serlin RC. Dimensions of the impact of cancer pain in a four country sample: new information from multidimensional scaling. Pain. 1996; 67(2-3):267-73.
10. Dorans NJ, Pommerich M, Holland PW. Linking and aligning scores and scales. New York: Springer; 2007.
11. Dorans NJ, Holland PW. Population invariance and the equatability of tests: Basic theory and the linear case. J Educ Meas. 2000; 37(4):281-306.
12. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 1999; 6(1):1-55.
13. Thissen D, Nelson L, Swygert KA. Item response theory applied to combinations of multiple chose and constructed response items - approximation methods for scale scores. In: Thissen D, Wainer H, editors. Test Scoring. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 2001. p. 293-341.
14. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. J Clin Epidemiol. 2011; 64(5):507-16.