Published in final edited form as:
J Anxiety Disord. 2014 January ; 28(1): 88–96. doi:10.1016/j.janxdis.2013.11.006.

Establishing a Common Metric for Self-Reported Anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety

Benjamin D. Schalet, Ph.D.[1], **Karon F. Cook, Ph.D.**[1], **Seung W. Choi, Ph.D.**[2], and **David Cella, Ph.D.**[1]

Benjamin D. Schalet: b-schalet@northwestern.edu; Karon F. Cook: karon.cook@northwestern.edu; Seung W. Choi: seung_choi@ctb.com; David Cella: d-cella@northwestern.edu

[1]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 625 Michigan Ave, 27th Floor, Chicago, IL, 60611

[2]CTB/McGraw-Hill, 20 Ryan Ranch Rd., Monterey, CA, 93940

Abstract

Researchers and clinicians wishing to assess anxiety must choose from among numerous assessment options, many of which purport to measure the same or a similar construct. A common reporting metric would have great value, and can be achieved when similar instruments are administered to a single sample and then linked to each other to produce cross-walk score tables. Using item response theory (IRT), we produced cross-walk tables linking three popular "legacy" anxiety instruments – MASQ (*N* = 743), GAD-7 (*N* = 748), and PANAS (*N* = 1120) – to the anxiety metric of the NIH Patient Reported Outcomes Measurement Information System (PROMIS®). The linking relationships were evaluated by resampling small subsets and estimating confidence intervals for the differences between the observed and linked PROMIS scores. Our results allow clinical researchers to retrofit existing data of three commonly-used anxiety measures to the PROMIS Anxiety metric and to compare clinical cut-off scores.

Keywords

anxiety; linking; PROMIS; MASQ; GAD-7; PANAS

1. Introduction

Researchers and clinicians wishing to assess anxiety in a clinical or community population must choose from among numerous assessment options, many of which purport to measure the same or a similar construct (Roemer, 2002; Harrington & Antony, 2008). A recent investigation found 92 empirically-based anxiety questionnaires (McHugh, Rasmussen, &

Corresponding author: Benjamin D. Schalet, Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 625 Michigan Ave, 27th Floor, Chicago, IL, 60611. b-schalet@northwestern.edu. Phone: +1-312-503-3640. Fax: +1-312-503-4800.

Otto, 2012). In choosing a questionnaire, users need to evaluate a number of issues, including the reported reliability and validity estimates of the instrument, the reading level required, the cost of the instrument, and whether the length of the instrument would unduly burden the patient/participant. Another important consideration is score comparability, i.e., whether a report of the scores will be useful to others in the field. That is, can the results obtained using instrument X in one set of studies be compared to results obtained using instrument Y in other studies? This concern may lead investigators to choose the most "popular" instrument, which may not always be the best instrument.

Self-report instruments typically are scored by summing or averaging individual item responses, leading to different score ranges for different instruments. This makes it difficult to compare the results across studies with different measures. Absent some method for aligning scores, one cannot know, for example, whether a mean summed MASQ score of 20 for a group of mental health outpatients falls above or below the case-defining score of 10 on the GAD-7 (Watson & Clark, 1999; Spitzer, Kroenke, Williams, & Löwe, 2006). One possible solution is to transform results to percentile rank scores or standardized scores, but this approach can be problematic, because these scores types are highly sensitive to sample characteristics such as restricted range (Baguley, 2009). If we convert our scores into percent of total maximum scores (Cohen, Cohen, Aiken, & West, 1999), the scores become independent of the particular sample, but not necessarily comparable across instruments, as the instruments may differ in their range of coverage.

The lack of standardized measurement among patient-reported outcome instruments was an impetus for the National Institutes of Health (NIH) Patient Reported Outcomes Measurement Information System (PROMIS®; Cella et al., 2010). Adapting the World Health Organization's (2007) tripartite framework of physical, mental, and social health, PROMIS researchers developed and calibrated multiple item banks (Buysse et al., 2010; Cella et al., 2010; Cella et al., 2007; Fries, Cella, Rose, Krishnan, & Bruce, 2009; Revicki et al., 2009), including one for measuring anxiety symptoms (Pilkonis et al., 2011). The PROMIS Anxiety bank -- comprising 29 items that include fear, anxious misery, and hyperarousal – can be administered as a brief computer adaptive test (CAT), an 8-item short form, or as an alternate subset of bank items that suit the investigator's needs (Cella et al., 2010; Pilkonis et al., 2011)

The Diagnostic and Statistical Manual of Mental Disorders – Fifth Edition (DSM-5) working groups have incorporated PROMIS items into their "review of systems" assessment (Narrow et al., 2013) and have recommend some PROMIS instruments as expanded modules (Kuhl, Kupfer, & Regier, 2011). As a result, it will be of interest to clinicians and researchers to document individual and grouped patient data in terms of PROMIS scores. However, some will continue to use instruments developed before PROMIS, and others will develop new instruments. Thus, there would be great value in having a common metric that associates PROMIS scores with scores from scales that measure the same or highly similar concepts (referred to hereafter as "legacy measures"). To create such a metric, we set out to "link" the scores from legacy measures to the PROMIS metric by establishing the mathematical relationships between legacy and PROMIS scores. Once scores are linked to a

common metric, a cross-walk table can be constructed that associates scores from one measure to corresponding scores on another.

The PROMIS metric uses the T-score, which is standardized with respect to mean (50) and standard deviation (10), centered around the US general population, matching the marginal distributions of gender, age, race, marital status, income, and education in the 2000 US census (Liu et al., 2010). Thus, a PROMIS Anxiety T-score of 60 can be interpreted as being one standard deviation higher (worse) than the "average person" in the US.

In this report, we present the results of studies linking scores of anxiety instruments under the PROsetta Stone® project. A detailed overview of linking, the PROsetta Stone methodology, and sample descriptions may be found elsewhere (Choi, Schalet, Cook, & Cella, 2013; see also Dorans, 2007). We linked scores from three legacy measures of general anxiety to the PROMIS Anxiety metric: the Mood and Anxiety Symptom Questionnaire (MASQ; Watson & Clark, 1991; Watson et al., 1995), the Generalized Anxiety Disorder Scale (GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006), and the Positive Affect and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988). Briefly, the PROsetta Stone methodology applies multiple linking methods, including those based in item response theory (IRT), as well as more traditional equipercentile methods (Lord, 1982). Such a multi-method approach is recommended by Kolen & Brennan (2004), as it serves to ensure that any violations of assumptions do not distort the results. Using a single group design (wherein each respondent answers questions on both the legacy and the PROMIS instrument), we test the accuracy of the each linking method by comparing the actual PROMIS scores with those obtained by linking. To evaluate bias and standard errors of the different linking methods, we apply a resampling analysis such that small subsets of cases (25, 50, and 75) are randomly drawn with replacement over 10,000 replications. For each replication, the mean difference between the actual and linked PROMIS score can be computed, allowing for an estimate of the confidence interval associated with linking, as a function of sample size (Choi, Schalet, Cook, & Cella, 2013).

2. Method

2.1. Measures

2.1.1. PROMIS Anxiety—The PROMIS Anxiety bank consists of 29 items with a 7-day time frame and a 5-point rating scale that ranges from 1 ("Never") to 5 ("Always") (Pilkonis et al., 2011, Cella et al., 2010). The item bank was developed using comprehensive mixed (qualitative and quantitative) methods (DeWalt, Rothrock, Yount, & Stone, 2007; Kelly et al., 2011), and focuses on fear (e.g., fearfulness, feelings of panic), anxious misery (e.g., worry, dread), hyperarousal (e.g., tension, nervousness, restlessness), and some somatic symptoms related to arousal (e.g., cardiovascular symptoms, dizziness). After confirming essential unidimensionality and fit to the graded response model (Samejima, 1969), one of a family of item response theory (IRT) models, items were calibrated with regard to their location (severity of anxiety) and discrimination (ability to distinguish people at different levels of anxiety). This produced a bank of questions that can accurately measure level of anxiety across its observed continuum, and provides the basis for innovative administration strategies such as computerized adaptive testing. The PROMIS Anxiety item bank provided

more information than conventional measures across a wider range of severity, ranging from normal to severely anxious (Pilkonis et al., 2011).

2.1.2. Mood and Anxiety Symptom Questionnaire (MASQ)—The MASQ is a 90-item self-report questionnaire that assesses depressive, anxious, and mixed symptomatology (Watson & Clark, 1991; Watson et al., 1995). The instrument requires that subjects to respond, on a Likert-type scale from 1 (“Not at All”) to 5 (“Extremely”), as to the presence and severity of a series of symptoms of anxiety. For the current analysis, we used the 11-item General Distress - Anxious Symptoms scale (MASQ-GA).

2.1.3. Generalized Anxiety Disorder Scale (GAD-7)—The GAD-7 is a 7-item instrument developed to identify likely cases of generalized anxiety disorder in primary care patients (Spitzer, Kroenke, Williams, & Löwe, 2006). Items are rated for the last two weeks, using a four-point rating scale for duration from 1 (“not at all”) to 5 (“nearly every day”). A score of 10 or greater on the GAD-7 represents a cut point for identifying cases of GAD, while cut points of 5, 10, and 15 might be interpreted as representing mild, moderate, and severe levels of anxiety on the GAD-7 (Spitzer, Kroenke, Williams, & Löwe, 2006).

2.1.4. Positive and Negative Affect Schedule (PANAS)—The PANAS is a 20-item instrument that comprises two scales measuring positive and negative affect, which may be described as general dimensions of mood (Watson, Clark, & Tellegen, 1988). Participants rate how they have felt over the past seven days using a five-point rating scale, ranging from 1 (“very slightly or not at all”) to 5 (“extremely”) on each of 20 words describing anxiety (“jittery”, “nervous”, “afraid”, and “scared”). For the current linking study, we used the 10-item negative affect scale (PANAS).

2.2. Samples

Our three linking samples were recruited from the US general population by internet panel survey providers. Demographic details for each sample are provided by Choi, Schalet, Cook, and Cella (2013).

The MASQ-GA linking sample included 743 individuals who were part of the original PROMIS calibration sample (Pilkonis et al., 2011). The MASQ-GA was administered along with the PROMIS Anxiety items. A detailed description of the study that finalized the PROMIS emotional distress measures is described elsewhere (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Pilkonis et al., 2011). The sample was 48% male and the mean age was 51 (*SD* = 19).

The GAD-7 linking sample included response to items collected in the NIH Toolbox study (HHSN260200600007; R. Gershon, PI). The NIH Toolbox initiative developed brief measures of emotional health, motor, cognitive, and sensory function. The PROMIS Anxiety measure is the basis for the NIH Toolbox Fear-Affect Survey, and the GAD-7 was the study’s legacy instrument also administered. Our GAD-7 linking sample was drawn from this dataset and included 748 participants who responded both to the PROMIS/Toolbox items and to the GAD-7 (Pilkonis et al., 2013). The sample was 44% male and the mean age was 47 (*SD* = 15).

The PANAS and PROMIS Anxiety response data were collected specifically for this linking study. Participants were recruited by Op4G, an internet survey company (www.op4g.com) that maintains a panel of respondents from the general population. A total of 1120 participants completed both measures. The sample was 47% male and the mean age was 46 (*SD* = 18).

In the MASQ-GA linking sample, all 29 PROMIS Anxiety items (Pilkonis et al., 2011) were administered. In the other two linking samples, subsets of the PROMIS Anxiety items were used: 20 items for the GAD-7 linking study and 15 items for the PANAS linking study. The PROMIS Anxiety item subsets were selected to be optimal in content coverage and measurement precision (Choi, Reise, Pilkonis, Hays, & Cella, 2010). Specifically, items were chosen to maximize the discrimination parameter, cumulatively represent a range of difficulty, and have diversity in item content. Both the 20-item and 15-item subsets had an estimated .99 correlation with the full 29-item bank (Pilkonis et al., 2013). Because the PROMIS Anxiety bank items are highly intercorrelated (mean adjusted item-total $r = .71$) and are sufficiently unidimensional, the principal difference between these sets (15 and 20 and 29 items) is their levels of precision, not content. For this reason, we did not limit our analysis to the 15 common PROMIS items included across all three data sets, but used the maximum number of PROMIS items available in each. Because PROMIS items are not scored as sums, but on a standardized T-score metric using IRT, scores obtained from different item subsets are directly comparable.

2.3. Analyses

Our analytic plan followed procedures reported in Choi, Schalet, Cook, and Cella (2013), which provides more detail on these methods. To meet the assumptions for linking, we first ensured that we were measuring essentially the same concept with PROMIS Anxiety and each legacy instrument (Dorans, 2007; Noonan et al., 2012). We did this by inspecting item content, calculating item-total correlations, conducting confirmatory factor analyses (CFAs), and estimating the proportion of general factor variance of the combined set items.

A second linking assumption is that the scores of the two measures to be linked are highly correlated. We calculated correlation coefficients between the raw scores of the linked measure and responses to the PROMIS Anxiety items. Following the recommendation of Dorans and Holland (2000), we tested a third linking assumption (subgroup invariance) by computing standardized Root Mean Square Deviation (RMSD). This statistic can be used to estimate the difference between the standardized difference of subpopulations (e.g., men and women) across two instruments. In all samples, we evaluated invariance for gender and age (over 65 / less than 65).

In addition to linking assumptions, we tested the unidimensionality assumption of IRT using both confirmatory and exploratory factor analytic methods. Since our planned IRT calibrations require only that the combined item set is essentially unidimensional, we conducted these analyses only on the combined items (e.g., PROMIS and the legacy measure). Confirmatory factor analyses (CFA) were conducted on the raw data treating the indicator variables as ordinal and using the WLSMV estimator of Mplus (Muthén & Muthén, 2006). This model posited that all items load highly on a single factor. Using

commonly used benchmark values (Lance, Butts, & Michels, 2006; Hopwood & Donnellan, 2006), model fit was evaluated based on standard fit indices including the Comparative Fit Index (CFI; > .90 adequate fit, > .95 very good fit), the Tucker Lewis Index (TLI; > .90 adequate fit, > .95 very good fit), and the Root Mean Square Error of Approximation (RMSEA; < .10 adequate fit, < .05 very good fit).

Next, we estimated the proportion of total variance attributable to a general factor (ω_h; McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005) using the **psych** package (Revelle, 2013) in **R** (R Core Development Team, 2011). This method estimates ω_h from the general factor loadings derived from an exploratory factor analysis and a Schmid–Leiman transformation (Schmid & Leiman, 1957). Values of .70 or higher for ω_h suggest that the item set is sufficiently unidimensional for most purposes (Reise, Scheines, Widaman, & Haviland, 2012).

We used two IRT-based and one non-IRT-based approach in linking the scores of measures. All item calibrations were based on the graded response model (Samejima, 1969). Both IRT-based approaches incorporated the established PROMIS calibrations (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Liu et al., 2010).

2.3.1. Fixed-Parameter Calibration—For each linking sample, item responses from a single legacy Anxiety measure (i.e., MASQ-GA, GAD-7, or PANAS) and the PROMIS Anxiety items were calibrated in a single run with PROMIS Anxiety item parameters fixed at their previously published values (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Pilkonis et al., 2011). The item parameters of the legacy Anxiety measures were estimated, subject to the metric defined by the PROMIS item parameters. Thus, this calibration yielded item parameters for the legacy measure that were on the PROMIS metric.

2.3.2. Separate Calibration with Linking Constants—The second IRT-based method we applied was separate calibration followed by the computation of transformation constants. This procedure uses the discrepancy between the established PROMIS parameters (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Pilkonis et al., 2011) and freely calibrated estimation of PROMIS parameters to place the legacy parameters on the established PROMIS metric. This is useful, because it avoids imposing the constraints inherent in the fixed-parameter calibration. We applied four procedures to obtain the linking constants: mean/mean, mean/sigma, Haebara (1980), and Stocking-Lord (1983). These IRT linking methods were implemented using the package **plink** (Weeks, 2010) in **R** (R Core Development Team, 2011). We ran all IRT calibrations using MULTILOG 7.03 (Thissen, Chen, & Bock, 2003).

2.3.3. Comparing IRT Linking Methods—We obtained four sets of IRT parameters for each legacy measure. To compare methods, we examined the differences between the test characteristic curves (TCCs). If the differences between the expected raw summed score values were small (e.g., less than 1 raw score point), we considered the methods interchangeable. When this occurred, we defaulted to the simpler fixed-parameter method for obtaining scores for each participant.

2.3.4. Equipercentile Linking—In each linking sample, we calculated scores on both the linked measure and the PROMIS Anxiety measure, along with each score's percentile rank within the sample. The scores of the two measures then were aligned by associating scores with equivalent percentile ranks on the two score distributions. The equipercentile linking was conducted to derive an equipercentile function using the LEGS program (Brennan, 2004). By applying the LEGS cubic-spline smoothing algorithm (Reinsch, 1967), the impact of random sampling error was minimized (Albano, 2011; Brennan, 2004; Kolen & Brennan, 2004).

2.3.5. Evaluation of Linking Methods—After applying all of the linking methods, we had multiple estimates of what a respondent's PROMIS scores would be based on their scores on a legacy measure. In addition, we had the person's actual score on the PROMIS Anxiety measure. We evaluated the accuracy of each linking approach by comparing respondents' linked scores to their actual scores on the PROMIS Anxiety metric. For each method, we computed correlations, and the mean and standard deviation of the difference in scores. To evaluate the bias and standard error of the different linking methods, we applied a resampling analysis such that small subsets of cases (25, 50, and 75) were randomly drawn with replacement over 10,000 replications. For each replication, the mean difference between the actual and linked PROMIS Anxiety T-score was computed. Then the mean and the standard deviation of the means were computed over replications as bias and empirical standard error, respectively. We then chose the most accurate linking method as a basis for the legacy-to-PROMIS cross-walk table.

3. Results

3.1. Item Content Overlap

Inspection of item content revealed substantial overlap between the PROMIS and legacy measures.On the MASQ-GA, seven out of ten items described feelings of fear, unease, and tension that corresponded to the content of PROMIS items. However, three items on the MASQ-GA describe specific somatic symptoms (diarrhea, lump in throat, upset stomach) not included in the PROMIS item banks. For the GAD-7, six items clearly correspond to the content coverage of PROMIS items; one item (irritability), however, does not overlap with a PROMIS Anxiety item. Six out of 10 items on the PANAS negative affect scale (afraid, scared, nervous, jittery, upset, and distressed) are very similar to items in the PROMIS Anxiety bank. Four items, however, describe different negative emotions (irritable, hostile, guilty, and ashamed).

3.2. Subpopulation Invariance

RMSD values for gender-related and age-related differences in these samples were uniformly low, such that less than 5.0% of the total variance could be explained by differences in subpopulations across instrument. These values meet the target (< 8%) recommended by Dorans and Holland (2000) to support subgroup invariance.

3.3. Classical Item Statistics

The classical item statistics on separate and combined instruments suggested relatively high levels of internal consistency and homogeneity to justify concordances between the PROMIS and each legacy measure (Table 1). Cronbach's internal consistency coefficients were high, ranging from .89 to .98 for the individual scales and .98 for each combined item set. Items in each of the three combined sets also were highly inter-correlated, with the mean adjusted item-total correlations ranging from .70 to .82. Correlations and disattenuated correlations (reported in parentheses) between scores on PROMIS Anxiety and the legacy scales were also high: 0.85 (.91) for the MASQ-GA, .86 (.91) for the GAD-7, and .89 (.93) for the PANAS. These values were, in one case, just below Dorans' recommended threshold for linking ($r \geq .86$), and met or exceeded the threshold in the other two cases (Dorans, 2004).

3.4. Unidimensionality of Combined Item Sets

For the 3 combined item sets of PROMIS and legacy items, CFA fit statistics ranged from adequate to very good, depending on the fit statistic referenced. PROMIS and MASQ-GA (48 items) fit values were: CFI = 0.951, TLI = 0.948, and RMSEA = 0.093. PROMIS and the GAD-7 (27 items) were CFI = 0.972, TLI = 0.970, and RMSEA = 0.077. For PROMIS and the PANAS (29 items) fit values were: CFI = 0.975, TLI = 0.972, and RMSEA = 0.102. These results suggest essential unidimensional data-model fit. Values of ω_h estimates were uniformly high: .90 (PROMIS and MASQ-GA), .95 (PROMIS and GAD-7), and .86 (PROMIS and PANAS). These values suggest the presence of a dominant general factor for each instrument pair (Reise, Scheines, Widaman, & Haviland, 2012).

3.5. IRT-based Linking

Table 2 displays the legacy instrument item parameters obtained from the fixed-parameter calibrations. For each instrument pair, the test characteristic curves (TCCs) of the separate calibrations using linking constants were nearly identical to the TCCs of the fixed calibrations. In fact, for each comparison between the TCCs, the expected raw summed score value differed by less than 1 point across thetas ranging from −4 to 4. Because of the close similarity of the different IRT solutions, we report only the results of the fixed-parameter estimates.

3.6. Equipercentile Linking

We mapped raw summed scores on the legacy instrument to raw summed scores on the PROMIS instrument. These score equivalents were then mapped to their corresponding PROMIS T-scores based on a raw-to-scale score conversion table. Because the raw summed score equivalents may take fractional values, such a conversion table was interpolated using statistical procedures (e.g., cubic spline; Brennan, 2004). Figure 1 shows the resulting equipercentile linking functions (in red) and the IRT cross-walk function (in black) for each linking pair. The two equipercentile functions shown incorporate post-smoothing values of 0.0 (no smoothing) and 1.0 (large smoothing; see Brennan, 2004). As Figure 1 shows, the scores derived from each of the methods were very similar, with the clear exception of MASQ-GA values greater than 43.

3.7. Evaluation of Linking Methods

To compare the accuracy of our linking methods, we compared the linked PROMIS T-score to the actual PROMIS T-score we obtained. We did this by computing the correlation, mean difference, and the standard deviation of difference scores for the linked and actual scores (see Table 3). The method labeled “IRT pattern scoring” refers to IRT scoring based on item parameter estimates and the pattern of responses to those items. The alternative, IRT summed or “cross-walk” scoring, also uses IRT. In this approach, however, the multiple response patterns that can result in the same summed score are assigned to the same scaled score. (This calculation is also used to construct the cross-walk tables [Appendix A].) Table 3 shows that IRT pattern scoring produced the best results for the MASQ-GA link; the correlation between actual and linked PROMIS T-scores was highest and the standard deviation of differences was lowest (mean differences are misleading because of negative and positive differences). For the GAD-7 and the PANAS, there were only very slight differences across the methods in terms of their comparison with the actual PROMIS scores obtained.

Results of the resampling technique with small subsets (n = 25; 50; 75) showed a similar picture in terms of method accuracy. Not surprisingly, as sample size increased from 25 to 75, the empirical standard error decreased. For the MASQ-GA link, the IRT pattern scoring showed the lowest empirical standard error (.59 for n = 75), followed by IRT cross-walk scoring (.64 for n = 75). For the GAD-7 link, the IRT and equipercentile methods produced virtually identical empirical standard errors (.72–.73 for n = 75), as did the PANAS link (.58–.60 for n = 75).

These standard errors can be used to create confidence intervals around linking results. If the PROsetta Stone cross-walk table was used to estimate PROMIS scores from a sample of 75 GAD-7 scores, there would be a 95% probability that the difference between the mean of this linked PROMIS score and the mean of the PROMIS Anxiety T-score (if obtained) would be within ± 1.41 T-score units (i.e., 1.96 × the 0.72 standard error for GAD-7).

3.8. Cross-walk Tables

Given that the IRT (fixed-calibration) linking method was as accurate or slightly more accurate than equipercentile methods, we used the item parameter estimates derived from the fixed-parameter calibration (see Table 2) to construct a cross-walk table by applying expected a posteriori (EAP) summed scoring. The tables for the MASQ-GA, GAD-7, and PANAS in Appendix A can be used to map simple raw summed scores from each legacy instrument to T-score values on the PROMIS Anxiety metric. Each raw summed score and corresponding PROMIS T-score is presented with the standard error associated with the scaled score. In place of the cross-walk table, researchers may wish to score their MASQ-GA, GAD-7, PANAS data on the PROMIS metric using the parameters in Table 2. IRT-pattern scoring is more accurate, and follows the standard PROMIS scoring procedure at www.assessmentcenter.net (Gershon et al., 2010).

3.9. Comparing Clinical Mean Scores

Figure 2 displays the linking functions for MASQ-GA, GAD-7, and PANAS that map their raw summed scores (the vertical axis) to the PROMIS Anxiety metric (the horizontal axis). For the GAD-7, a score of 10 represents the case-defining cut-score; Spitzer, Kroenke, Williams, & Löwe (2006) found that 89% of patients diagnosed with generalized anxiety disorder had a score of 10 or higher, while 82% of those without the disorder scored less than 10. In our data, this score is linked to PROMIS Anxiety T-score of 62.3. This value is close to the tentative threshold PROMIS investigators have set on the anxiety measure of 60, or one standard deviation above the population mean (Cella et al., 2008).

Although there are no established norms or cut-offs for the MASQ-GA and PANAS scales, we briefly compare the mean scores of reported clinical samples to the equivalent linked PROMIS scores. Watson and Clark (1999) report PANAS negative affect means for a psychiatric inpatient sample (M= 25.5) and a mixed clinical sample (M= 26.3) using the “in general” time context. Using a PANAS score of 26 as a provisional cut-score, we find that it is linked to a PROMIS score of 61.7. For the MASQ-GA, young people seeking outpatient treatment showed a mean score of 25.2 (Buckby, Yung, Cosgrave, & Killackey, 2007), while a sample of mood and anxiety disorder patients scored an average of 24.8 (Boschen & Oei, 2007). An MASQ-GA score of 25 is linked to 60.1 on the PROMIS anxiety scale.

4. Discussion

This paper represents the first effort to link multiple measures of anxiety to the PROMIS metric. Although investigators of patient-reported outcomes have linked measures in a number of domains -- including depression (Choi, Schalet, Cook, & Cella, 2013; Fischer, Tritt, Klapp, & Fliege, 2011), fatigue (Noonan et al., 2011; Holzner et al., 2006), and pain (Chen, Revicki, Lai, Cook, & Amtmann, 2009) -- we are unaware of any study that has linked anxiety instruments. This work has resulted in several useful products: three cross-walk tables, as well as item parameters for legacy measures linked to the PROMIS Anxiety metric (for converting legacy scores to the PROMIS Anxiety metric using pattern scoring).

Our study has a number of prominent strengths. First, it follows a single-group design, which produces the most robust links (Dorans, 2007). The single-group design also allowed us to measure the accuracy of the linking by examining differences between the actual score and the one predicted by the linking. Second, we employed multiple linking methods so that we could empirically determine which (if any) method minimized differences between observed and linked scores. Third, our calibrations were not determined by the current sample, but were anchored on the PROMIS calibrations that were derived from the larger standardization sample (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Pilkonis et al., 2011) and centered on the 2000 US Census (Liu et al., 2010).

Our results also have implications for clinical practice and psychopathology research. As Figure 2 illustrates, the clinical mean scores and published cut-off scores converged around one standard deviation above the US population mean (PROMIS T-scores of 60.1 – 62.4). Interestingly, one standard deviation above the mean corresponds to percentile ranks of 82 (women) and 88 (men) (Gershon et al., 2010), which is within the range of estimated 12-

month US prevalence rates (11–18%) based on diagnostic criteria for any anxiety disorders (Kessler et al., 2006; Grant et al., 2004; Kessler et al., 1994). The similarity among these T-scores and prevalence rates suggests that self-report questionnaires of severity can detect non-specific anxiety symptoms that are clinically significant and warrant more precise diagnosis.

Watson and colleagues proposed a hierarchical model that differentiates general and specific symptoms among “emotional disorders” (e.g., Watson, O’Hara, & Stuart, 2008). In this conceptualization, PROMIS Anxiety and our three legacy measures are best understood as measures of general anxious distress. Users seeking to assess narrower symptom groups, such as specific phobias or somatic arousal, are advised to select more targeted measures, such as the NIH Toolbox Somatic Arousal scale (Pilkonis et al., 2013; see also McHugh, Rasmussen, & Otto, 2011) or diagnostic instruments. Notably, the few somatic symptoms included in the MASQ-GA also show relatively low discrimination values on the PROMIS metric (Table 2). These values ranged from 0.79 to 1.25, corresponding to factor loadings of .42 to .59 (Wirth & Edwards, 2007). Extending assessments to include somatic symptoms would facilitate discrimination among the emotional disorders (Aldao, Mennin, Linardatos, & Fresco, 2010).

Researchers and clinicians interested in linking any of the three legacy measures to PROMIS Anxiety have three options. First, they can use the cross-walk chart to substitute each participant’s summed legacy score with the corresponding PROMIS T-score. The scores then can be used for descriptive and inferential analyses. Second, researchers can use the item parameter estimates we obtained for the legacy measures and IRT software (e.g., IRTPRO (Cai, Thissen, du Toit, 2011) or Firestar (Choi, 2009)) to obtain scores based on participants’ responses to the items. This approach in general will yield slightly more accurate results than the cross-walk table and also has the advantage of accounting for missing data without explicit imputation. Finally, summary (not individual) sample scores from legacy measures (e.g., as in the case of meta-analyzing published research) can be cross-walked to PROMIS scores.

Despite the strengths of our linking methodology, scores linked to the PROMIS metric based on legacy scores may have more error than scores obtained directly from the PROMIS Anxiety measure (i.e., linking error is added to measurement error). In addition, the standard errors for using the cross-walk tables with samples of less than 75 participants will increase and may not be adequate for some purposes. For example, the standard error for linking GAD-7 scores increases from .72 for a sample of 75 to 6.56 for a sample of 1 participant. Finally, our linking tables should be used with the knowledge that concordances between any two instruments (regardless of statistical method) may be sensitive to population differences (Dorans, 2007).

In addition, although we reported on a resampling analysis, it would have been ideal to evaluate the robustness of the linking relationship on a new sample (independent of the sample from which the linking relationship was derived). Such a sample would be used to examine empirically the bias and standard error of the linking results. The small confidence

interval we constructed using the resampling technique (e.g., ± 1.41 T-score units for GAD-7 linking, $n = 75$) may underestimate the error introduced by the linking procedure.

In conclusion, this is the first report in health measurement that links multiple legacy scale to the PROMIS Anxiety metric. We provided several tools for researchers to link scores from three popular anxiety measures to the PROMIS Anxiety metric. Research is underway to complete a large number of additional linking studies to establish a mathematical bridge from scores on legacy instruments measuring other domains to the PROMIS metric.

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, please see www.prosettastone.org. We would like to thank Joshua Rutsohn and Helena Correia for their help in the preparation of this manuscript.

References

Albano, T. equate: Statistical Methods for Test Score Equating (R Package Version 1.1–4) [Computer software]. 2011. Retrieved from http://cran.opensourceresources.org/web/packages/equate/equate.pdf

Aldao AA, Mennin DS, Linardatos EE, Fresco DM. Differential patterns of physical symptoms and subjective processes in generalized anxiety disorder and unipolar depression. Journal of Anxiety Disorders. 2010; 24(2):250–259.10.1016/j.janxdis.2009.12.001 [PubMed: 20060680]

Baguley T. Standardized or simple effect size: What should be reported? British Journal of Psychology. 2009; 100(3):603–617.10.1348/000712608X377117 [PubMed: 19017432]

Brennan, R. Linking with Equivalent Group or Single Group Design (LEGS) (Version 2.0) [Computer software]. Iowa City, IA: University of Iowa: Center for Advanced Studies in Measurement and Assessment (CASMA); 2004.

Boschen MJ, Oei TP. Discriminant validity of the MASQ in a clinical sample. Psychiatry Research. 2007; 150(2):163–171.10.1016/j.psychres.2006.03.008 [PubMed: 17292971]

Buckby JA, Yung AR, Cosgrave EM, Killackey EJ. Clinical utility of the Mood and Anxiety Symptom Questionnaire (MASQ-GA) in a sample of young help-seekers. BMC Psychiatry. 2007; 710.1186/1471-244X-7-50

Buysse DJ, Yu L, Moul DE, Germain A, Stover A, Dodds NE, Pilkonis PA. Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. Sleep. 2010; 33(6):781–792. [PubMed: 20550019]

Cai, L.; Thissen, D.; du Toit, S. IRTPRO 2.01. Lincolnwood, IL: Scientific Software International; 2011.

Cella D, Choi S, Rosenbloom S, Surges Tatum D, Garcia S, Lai J-S, George J, Gershon R. A novel IRT-based case-ranking approach to derive expert standards for symptom severity. Quality of Life Research, ISOQOL Conference Supplement. 2008:A-32.

Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Hays R. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology. 2010; 63(11): 1179–1194.10.1016/j.jclinepi.2010.04.011 [PubMed: 20685078]

Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. Medical Care. 2007; 45(5, Suppl1):S3–S11.10.1097/01.mlr.0000258615.42478.55 [PubMed: 17443116]

Chen W, Revicki DA, Lai J, Cook KF, Amtmann D. Linking Pain Items from Two Studies Onto a Common Scale Using Item Response Theory. Journal of Pain & Symptom Management. 2009; 38(4):615–628.10.1016/j.jpainsymman.2008.11.016 [PubMed: 19577422]

Choi SW. Firestar: Computerized adaptive testing simulation program for polytomous IRT models. Applied Psychological Measurement. 2009; 33(8):644–645.10.1177/0146621608329892

Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. Quality of Life Research. 2010; 19:125–136.10.1007/s11136-009-9560-5 [PubMed: 19941077]

Choi SW, Schalet BD, Cook KF, Cella D. Establishing a Common Metric for Depressive Symptoms: Linking BDI-II, CES-D, and PHQ-9 to PROMIS Depression. 2013 Manuscript submitted for publication.

Cohen P, Cohen J, Aiken LS, West SG. The problem of units and the circumstance for POMP. Multivariate Behavioral Research. 1999; 34(3):315–346.10.1207/S15327906MBR3403_2

DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: The PROMIS qualitative item review. Medical Care. 2007; 45(Suppl 1):S12–S21.10.1097/01.mlr.0000254567.79743.e2 [PubMed: 17443114]

Dorans NJ. Equating, Concordance, and Expectation. Applied Psychological Measurement. 2004; 28(4):227–246.10.1177/014662160426503

Dorans NJ. Linking scores from multiple health outcome instruments. Quality of Life Research. 2007; 16(Suppl1):85–94.10.1007/s11136-006-9155-3 [PubMed: 17286198]

Dorans NJ, Holland PW. Population invariance and the equatability of tests: Basic theory and the linear case. Journal of Educational Measurement. 2000; 37(4):281–306.10.1111/j.1745-3984.2000.tb01088.x

Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. Journal of Rheumatology. 2009; 36(9):2061–2066.10.3899/jrheum.090358 [PubMed: 19738214]

Fischer H, Tritt K, Klapp BF, Fliege H. How to compare scores from different depression scales: Equating the Patient Health Questionnaire (PHQ) and the ICD-10- Symptom Rating (ISR) using item response theory. International Journal of Methods in Psychiatric Research. 2011; 20(4):203–214.10.1002/mpr.350 [PubMed: 22021205]

Gershon RC, Rothrock N, Hanrahan R, Bass M, Cella D. The use of PROMIS and Assessment Center to deliver patient-reported outcome measures in clinical research. Journal of Applied Measurement. 2010; 11:304–314. [PubMed: 20847477]

Haebara T. Equating logistic ability scales by a weighted least squares method. Japanese Psychological Research. 1980; 22(3):144–149.

Harrington JL, Antony MM. Assessment of Anxiety Disorders. Oxford Handbook of Anxiety and Related Disorders. 2008:277.

Holzner B, Bode RK, Hahn EA, Cella D, Kopp M, Sperner-Unterweger B, Kemmler G. Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research. European Journal of Cancer. 2006; 42(18):3169–3177. [PubMed: 17045472]

Hopwood CJ, Donnellan M. How should the internal structure of personality inventories be evaluated? Personality and Social Psychology Review. 2010; 14(3):332–346.10.1177/1088868310361240 [PubMed: 20435808]

Kelly MAR, Morse JQ, Stover A, Hofkens T, Huisman E, Shulman S, Pilkonis PA. Describing depression: Congruence between patient experiences and clinical assessments. British Journal of Clinical Psychology. 2011; 50:46–66.10.1348/014466510X493926 [PubMed: 21332520]

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

Kuhl EA, Kupfer DJ, Regier DA. Patient-centered revisions to the DSM-5. Virtual Mentor. 2011; 13(12):873. Retrieved from http://virtualmentor.ama-assn.org/2011/12/pdf/stas1-1112.pdf. [PubMed: 23137425]

Liu HH, Cella D, Gershon R, Shen J, Morales LS, Riley W, Hays RD. Representativeness of the PROMIS Internet panel. Journal of Clinical Epidemiology. 2010; 63 (11):1169–78.10.1016/j.jclinepi.2009.11.021 [PubMed: 20688473]

Lance C, Butts M, Michels L. The sources of four commonly reported cutoff criteria: What did they really say? Organizational ResearchMethods. 2006; 9:202–220.10.1177/1094428105284919

Lord FM. The Standard Error of Equipercentile Equating. Journal of Educational and Behavioral Statistics. 1982; 7(3):165–174.10.3102/10769986007003165

McDonald, RP. Test Theory: A Unified Treatment. Psychology Press; 1999.

McHugh RK, Rasmussen JL, Otto MW. Comprehension of self-report evidence-based measures of anxiety. Depression and Anxiety. 2011; 28(7):607–614.10.1002/da.20827 [PubMed: 21618668]

Muthén, LK.; Muthén, BO. Mplus. [Computer software]. Los Angeles: Muthén & Muthén; 2006.

Narrow WE, Clarke DE, Kuramoto SJ, Kraemer HC, Kupfer DJ, Greiner L, Regier DA. DSM-5 Field Trials in the United States and Canada, part III: Development and reliability testing of a cross-cutting symptom assessment for DSM-5. American Journal of Psychiatry. 2013; 170(1):71–82.10.1176/appi.ajp.2012.12071000 [PubMed: 23111499]

Noonan VK, Cook KF, Bamer AM, Choi SW, Kim J, Amtmann D. Measuring fatigue in persons with multiple sclerosis: Creating a crosswalk between the Modified Fatigue Impact Scale and the PROMIS Fatigue Short Form. Quality of Life Research. 2012; 21(7):1123–1133.10.1007/s11136-011-0040-3 [PubMed: 22048931]

Pilkonis PA, Choi SW, Salsman JM, Butt Z, Moore TL, Lawrence SM, Cella D. Assessment of self-reported negative affect in the NIH Toolbox. Psychiatry Research. 2013; 206(1):88–97.10.1016/j.psychres.2012.09.034 [PubMed: 23083918]

Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. Assessment. 2011; 18(3):263–283.10.1177/1073191111411667 [PubMed: 21697139]

R Core Development Team. R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing; 2011. Retrieved from http://www.r-project.org/

Reinsch CH. Smoothing by spline functions. Numerische Mathematik. 1967; 10(3):177–183.10.1007/BF02162161

Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. Educational and Psychological Measurement. 2012; 73(1):5–26.10.1177/0013164412449831

Revicki DA, Chen W, Harnam N, Cook KF, Amtmann D, Callahan LF, Keefe FJ. Development and psychometric analysis of the PROMIS pain behavior item bank. Pain. 2009; 146(1–2):158–169.10.1016/j.pain.2009.07.029 [PubMed: 19683873]

Revelle, W. psych: Procedures for personality and psychological research (R Package Version 1.2.8) [Computer software]. Evanston, IL: Northwestern University; 2013. Retrieved from http://cran.r-project.org/web/packages/psych/index.html

Roemer, L. Measures for anxiety and related constructs. In: Antony, MM.; Orsillo, SM.; Roemer, L., editors. Practitioner's guide to empirically based measures of anxiety. New York, US: Kluwer Academic Publishers; 2002. p. 49-83.

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph. 1969:17.10.1007/BF02290599

Schmid JJ, Leiman JM. The development of hierarchical factor solutions. Psychometrika. 1957; 22:53–61.10.1007/BF02289209

Spitzer RL, Kroenke K, Williams JB, Lowe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Archives of Internal Medicine. 2006; 166(10):1092.10.1001/archinte.166.10.1092 [PubMed: 16717171]

Stocking ML, Lord FM. Developing a common metric in item response theory. Applied Psychological Measurement. 1983; 7:201–210.10.1177/014662168300700208

Thissen, D.; Chen, W-H.; Bock, RD. Multilog 7.03 [Computer software]. Lincolnwood, IL: Scientific Software International; 2003.

Watson D, Clark LA, Tellegen A. Development and validation of brief measures of Positive and Negative Affect: The PANAS Scales. Journal of Personality and Social Psychology. 1988; 54:1063–1070.10.1037/0022-3514.54.6.1063 [PubMed: 3397865]

Watson, D.; Clark, LA. Unpublished manuscript. University of Iowa, Department of Psychology; Iowa City: 1991. The Mood and Anxiety Symptom Questionnaire.

Watson D, Clark LA, Weber K, Smith Assenheimer J, Strauss ME, McCormick RA. Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. Journal of Abnormal Psychology. 1995; 104:15–15. [PubMed: 7897037]

Watson, D.; Clark, LA. The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form. Unpublished manuscript. 1999. Retrieved from http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1011&context=psychology_pubs

Watson D, O'Hara MW, Stuart S. Hierarchical structures of affect and psychopathology and their implications for the classification of emotional disorders. Depression and Anxiety. 2008; 25(4): 282–288.10.1002/da.20496 [PubMed: 18415951]

Weeks JP. Plink: An R package for linking mixed-format tests using IRT-based methods. Journal of Statistical Software. 2010; 35(12):1–33. [PubMed: 21603108]

Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. Psychological Methods. 2007; 12(1):58.10.1037/1082-989X.12.1.58 [PubMed: 17402812]

World Health Organization. Constitution of the World Health Organization: Basic documents. 46. Geneva,Switzerland: Author; 2007.

Zinbarg RE, Revelle W, Yovel I, Li W. Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika. 2005; 70:123–133.10.1007/s11336-003-0974-7

Appendix A

Table A.1

Raw Score to T-Score Conversion Table (IRT Fixed-Parameter Calibration Linking) for MASQ-GA to PROMIS Anxiety

MASQ-GA Score	PROMIS T-score	T-Score SE
11	35.2	6.1
12	38.9	5.4
13	41.9	5.0
14	44.3	4.6
15	46.5	4.3
16	48.4	4.0
17	50.1	3.8
18	51.7	3.6
19	53.1	3.5
20	54.3	3.4
21	55.6	3.4
22	56.7	3.3
23	57.9	3.3
24	59.0	3.3
25	60.1	3.2
26	61.1	3.2
27	62.1	3.2
28	63.2	3.1
29	64.1	3.1
30	65.1	3.1
31	66.1	3.1
32	67.1	3.1

MASQ-GA Score	PROMIS T-score	T-Score SE
33	68.0	3.1
34	69.0	3.2
35	69.9	3.2
36	70.9	3.2
37	71.9	3.2
38	72.9	3.3
39	73.9	3.3
40	74.9	3.4
41	76.0	3.5
42	77.1	3.5
43	78.2	3.6
44	79.3	3.6
45	80.3	3.6
46	81.4	3.6
47	82.4	3.5
48	83.3	3.4
49	84.2	3.3
50	85.0	3.1
51	85.7	2.9
52	86.4	2.6
53	86.9	2.4
54	87.4	2.2
55	87.7	1.9

Table A.2

Raw Score to T-Score Conversion Table (IRT Fixed-Parameter Calibration Linking) for GAD-7 to PROMIS Anxiety

GAD-7 Score	PROMIS T-score	T-Score SE
0	38.5	6.1
1	44.5	4.6
2	47.9	4.0
3	50.4	3.7
4	52.6	3.5
5	54.6	3.4
6	56.3	3.3
7	57.9	3.3
8	59.4	3.3
9	60.9	3.2
10	62.3	3.2
11	63.7	3.2
12	65.0	3.1

GAD-7 Score	PROMIS T-score	T-Score SE
13	66.4	3.1
14	67.7	3.1
15	69.0	3.1
16	70.4	3.2
17	71.9	3.3
18	73.5	3.4
19	75.3	3.6
20	77.2	3.7
21	80.1	4.1

Table A.3

Raw Score to T-Score Conversion Table (IRT Fixed-Parameter Calibration Linking) for PANAS to PROMIS Anxiety

PANAS Score	PROMIS T-score	T-score SE
10	37.4	5.9
11	43.0	4.2
12	46.0	3.7
13	48.1	3.3
14	49.9	3.0
15	51.3	2.8
16	52.6	2.7
17	53.8	2.6
18	54.8	2.5
19	55.8	2.4
20	56.7	2.4
21	57.6	2.4
22	58.5	2.3
23	59.3	2.3
24	60.1	2.3
25	60.9	2.3
26	61.7	2.3
27	62.4	2.3
28	63.2	2.3
29	63.9	2.3
30	64.7	2.3
31	65.5	2.3
32	66.2	2.3
33	67.0	2.3
34	67.7	2.3
35	68.5	2.3
36	69.3	2.3

PANAS Score	PROMIS T-score	T-score SE
37	70.1	2.3
38	70.9	2.3
39	71.8	2.3
40	72.6	2.3
41	73.5	2.4
42	74.5	2.4
43	75.5	2.4
44	76.5	2.5
45	77.6	2.5
46	78.8	2.6
47	80.2	2.7
48	81.7	2.8
49	83.4	2.9
50	85.1	2.8

Highlights

We produced cross-walk tables linking three popular instruments to PROMIS Anxiety.

The scores of our common measure (PROMIS Anxiety) are centered on the 2000 US Census.

Users can directly compare clinical scores obtained on multiple measures of anxiety.

Clinical means or cut-off scores were close to one SD above the population mean.

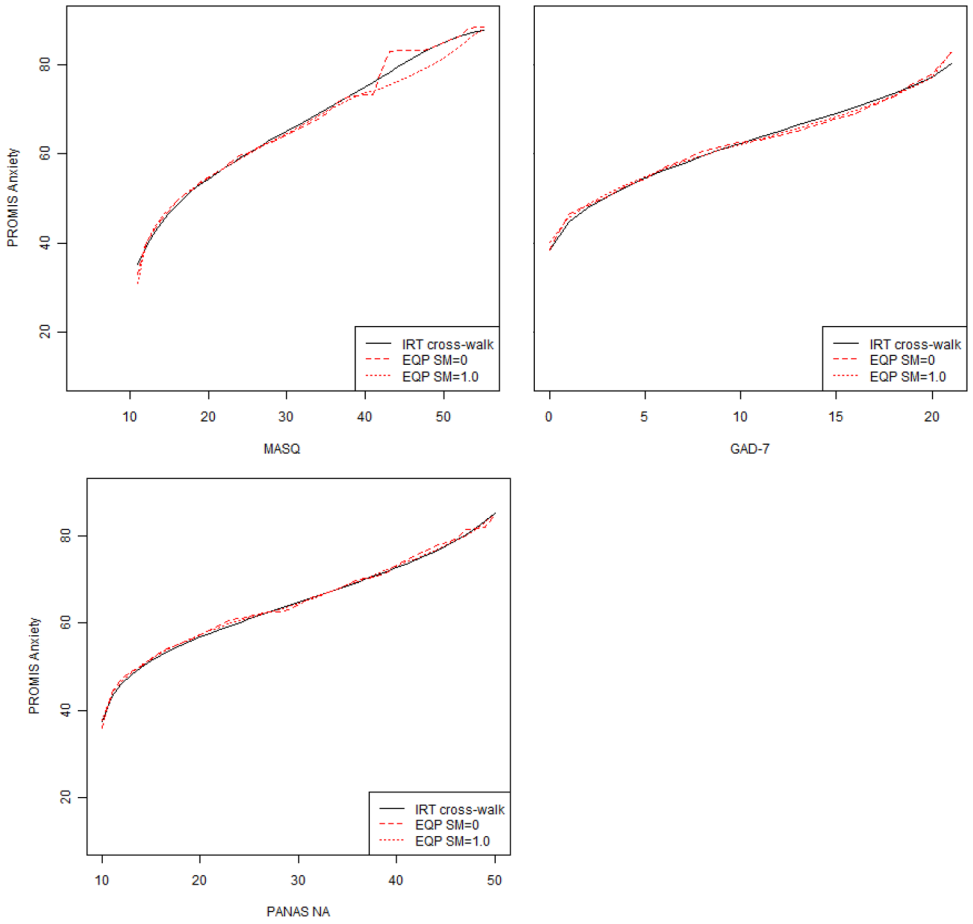

Figure 1.

Figure 1a–c. IRT cross-walk function (based on fixed-parameter calibration) and equipercentile functions with different levels of smoothing. EQP = Equipercentile; SM = Post-Smoothing.

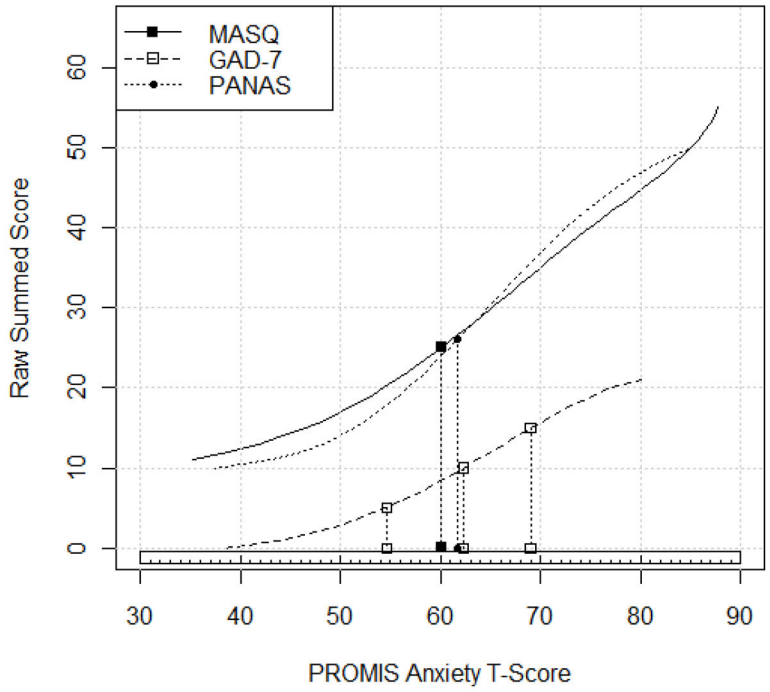

Figure 2.

Comparison of clinically relevant scores on the PROMIS Anxiety metric. MASQ-GA: 25 for the mean of a general clinical sample with mood and anxiety disorders (Boschen & Oei, 2007); GAD-7: 5, 10, and 15 correspond to mild, moderate, and severe symptoms of generalized anxiety disorder (Spitzer, Kroenke, Williams, & Löwe, 2006); PANAS: 26 for the mean of a clinically mixed sample (Watson & Clark, 1999).

Table 1

Classical Item Analysis

				Adjusted (corrected for overlap) Item-total Correlation		
	Number of Items	***N***	**Cronbach's Alpha Reliability**	**Minimum**	**Mean**	**Maximum**
PROMIS Anxiety	29	743	0.97	0.51	0.73	0.83
MASQ-GA	11	743	0.89	0.43	0.62	0.78
PROMIS & MASQ-GA	40	743	0.98	0.39	0.70	0.83
PROMIS Anxiety	20	748	0.97	0.61	0.79	0.88
GAD-7	7	748	0.93	0.71	0.78	0.85
PROMIS & GAD-7	27	748	0.98	0.60	0.78	0.87
PROMIS Anxiety	15	1120	0.98	0.79	0.84	0.88
PANAS	10	1120	0.95	0.74	0.80	0.84
PROMIS & PANAS	25	1120	0.98	0.70	0.82	0.86

Table 2

Transformed Item Parameter Estimates (Fixed-Parameter Calibration)

	MASQ-GA					GAD-7				PANAS				
Item	**Slope**	**CB1**	**CB2**	**CB3**	**CB4**	**Slope**	**CB1**	**CB2**	**CB3**	**Slope**	**CB1**	**CB2**	**CB3**	**CB4**
1	2.47	0.63	1.76	2.55	4.16	2.38	0.19	1.55	2.30	2.69	0.03	1.00	1.80	2.85
2	0.79	1.22	3.10	4.63	7.86	2.62	0.27	1.44	2.08	2.39	−0.16	0.95	1.78	2.99
3	2.98	−0.04	1.19	1.87	3.03	2.53	0.04	1.35	1.99	2.45	0.53	1.23	2.01	2.92
4	3.06	−0.12	1.05	1.77	2.63	2.21	0.11	1.26	1.95	2.86	0.60	1.33	1.97	2.76
5	1.25	1.61	3.22	3.79	4.75	1.98	0.78	1.91	2.73	1.82	0.62	1.56	2.38	3.50
6	1.09	0.25	2.00	3.19	5.49	1.66	0.19	1.70	2.59	2.01	−0.33	0.97	1.79	2.78
7	2.31	−0.12	1.02	1.75	2.75	2.26	0.68	1.81	2.39	2.48	0.71	1.35	1.97	2.72
8	2.18	−0.01	1.07	1.77	2.78					3.22	0.14	0.99	1.60	2.39
9	1.16	0.98	2.51	3.69	5.23					2.69	0.42	1.23	1.88	2.68
10	2.34	0.03	1.06	1.64	2.64					3.40	0.62	1.28	1.81	2.57
11	0.84	−1.11	1.00	2.74	4.83									

Table 3

Correlations, Mean Differences, and Standard Deviations of Actual vs. Linked PROMIS Anxiety T-Scores

Linking Method / Instrument	Correlation	Mean Difference	SD of Differences
MASQ-GA			
IRT Pattern Scoring	0.85	−0.07	5.36
IRT Cross-walk Scoring	0.82	0.02	5.86
EQP, SM=0.0	0.81	−0.10	6.05
EQP, SM=1.0	0.80	0.03	6.19
GAD-7			
IRT Pattern Scoring	0.83	0.16	6.56
IRT Cross-walk Scoring	0.82	0.24	6.59
EQP, SM=0.0	0.82	−0.01	6.61
EQP, SM=1.0	0.83	−0.55	6.48
PANAS			
IRT Pattern Scoring	0.89	0.10	5.17
IRT Cross-walk Scoring	0.89	0.26	5.26
EQP, SM=0.0	0.89	0.16	5.39
EQP, SM=1.0	0.89	−0.14	5.24

Note. EQP = Equipercentile; SM = Post-Smoothing.