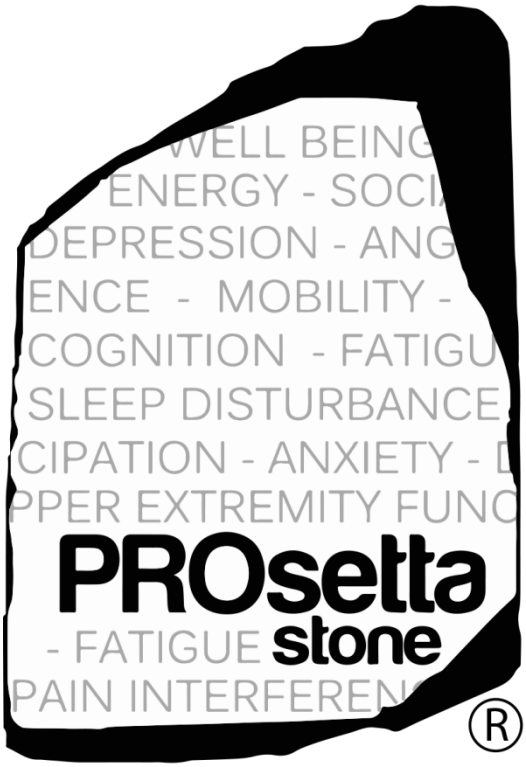

PROSETTA STONE® ANALYSIS REPORT

A ROSETTA STONE FOR PATIENT REPORTED OUTCOMES

PROMIS DEPRESSION AND PATIENT HEALTH QUESTIONNAIRE (PHQ-2)

DAVID CELLA , BENJAMIN D. SCHALET, MICHAEL KALLEN, JIN-SHEI LAI, KARON F. COOK, JOSHUA RUTSOHN & SEUNG W. CHOI

DEPARTMENT OF MEDICAL SOCIAL SCIENCES
FEINBERG SCHOOL OF MEDICINE
NORTHWESTERN UNIVERSITY

This research was supported by an NIH/National Cancer Institute grant PROSETTA STONE (1RC4CA157236-01, PI: David Cella). Authors acknowledge careful reviews, comments, and suggestions from Drs. Robert Brennan, Lawrence Hedges, Won-Chan Lee, and Nan Rothrock.

Table of Contents

PRO Rosetta Stone (*PROsetta Stone*®) Analysis

1. Introduction

A common problem when using a variety of patient-reported outcome measures (PROs) for diverse populations and subgroups is establishing the comparability of scales or units on which the outcomes are reported. The lack of comparability in metrics (e.g., raw summed scores vs. scaled scores) among different PROs poses practical challenges in measuring and comparing effects across different studies. Linking refers to establishing a relationship between scores on two different measures that are not necessarily designed to have the same content or target population. When tests are built in such a way that they differ in content or difficulty, linking must be conducted in order to establish a relationship between the test scores. One technique, commonly referred to as equating, involves the process of converting the system of units of one measure to that of another. This process of deriving equivalent scores has been used successfully in educational assessment to compare test scores obtained from parallel or alternate forms that measure the same characteristic with equal precision. Extending the technique further, comparable scores are sometimes derived for measures of different but related characteristics. The process of establishing comparable scores generally has little effect on the magnitude of association between the measures. Comparability may not signify interchangeability unless the association between the measures approaches the reliability. Equating, the strongest form of linking, can be established only when two tests 1) measure the same content/construct, 2) target very similar populations, 3) are administered under similar conditions such that the constructs measured are not differentially affected, 4) share common measurement goals and 5) are equally reliable. When test forms are created to be similar in content and difficulty, equating adjusts for differences in difficulty. Test forms are considered to be essentially the same, so scores on the two forms can be used interchangeably after equating has adjusted for differences in difficulty. For tests with lesser degrees of similarity, only weaker forms of linking are meaningful, such as calibration, concordance, projection, or moderation.

2. The PRO Rosetta Stone Project

The primary aim of the PRO Rosetta Stone (PROsetta Stone®) project (1RC4CA157236-01, PI: David Cella) is to develop and apply methods to link the Patient-Reported Outcomes Measurement Information System (PROMIS) measures with other related “legacy” instruments to expand the range of PRO assessment options within a common, standardized metric. The project identifies and applies appropriate linking methods that allow scores on a range of PRO instruments to be expressed as standardized T-score metrics linked to the PROMIS. This report (Volume 2) encompasses 23 linking studies based on available PRO data that are primarily from PROsetta Stone Waves 1 and 2, as well as a few links based on PROMIS Wave 1 and

NIH Toolbox. The PROsetta Stone Report Volume 1 included linking results primarily from PROMIS Wave 1, as well as links based on NIH Toolbox and Neuro-QoL data.

2.1 Patient-Reported Outcomes Measurement Information System (PROMIS)

In 2004, the NIH initiated the PROMIS[1] cooperative group under the NIH Roadmap[2] effort to re-engineer the clinical research enterprise. The aim of PROMIS is to revolutionize and standardize how PRO tools are selected and employed in clinical research. To accomplish this, a publicly-available system was developed to allow clinical researchers access to a common repository of items and state-of-the-science computer-based methods to administer the PROMIS measures. The PROMIS measures include item banks across a wide range of domains that comprise physical, mental, and social health for adults and children, with 12-124 items per bank. Initial concepts measured include emotional distress (anger, anxiety, and depression), physical function, fatigue, pain (quality, behavior, and interference), social function, sleep disturbance, and sleep-related impairment. The banks can be used to administer computerized adaptive tests (CAT) or fixed-length forms in these domains. We have also developed 4 to 20-item short forms for each domain, and a 10-item Global Health Scale that includes global ratings of five broad PROMIS domains and general health perceptions. As described in a full issue of *Medical Care* (Cella et al., 2007), the PROMIS items, banks, and short forms were developed using a standardized, rigorous methodology that began with constructing a consensus-based PROMIS domain framework.

All PROMIS banks have been calibrated according to Samejima (Samejima, 1969) graded response model (based on large data collections including both general and clinical samples) and re-scaled (mean=50 and SD=10) using scale-setting subsamples matching the marginal distributions of gender, age, race, and education in the 2000 US census. The PROMIS Wave I calibration data included a small number of full-bank testing cases (approximately 1,000 per bank) from a general population taking one full bank and a larger number of block-administration cases (n= ~14,000) from both general and clinical populations taking a collection of blocks representing all banks with 7 items each. The full-bank testing samples were randomly assigned to one of 7 different forms. Each form was composed of one or more PROMIS domains (with an exception of Physical Function where the bank was split over two forms) and one or more legacy measures of the same or related domains.

The PROMIS Wave I data collection design included a number of widely accepted “legacy” measures. The legacy measures used for validation evidence included Buss-Perry Aggression Questionnaire (BPAQ), Center for Epidemiological Studies Depression Scale (CES-D), Mood and Anxiety Symptom Questionnaire (MASQ), Functional Assessment of Chronic Illness

[1] www.nihpromis.org
[2] www.nihroadmap.nih.gov

Therapy-Fatigue (FACIT-F), Brief Pain Inventory (BPI), and SF-36. Furthermore, included within each of the PROMIS banks were items from several other existing measures. Depending on the nature and strength of relationship between the measures, various linking procedures can be used to allow for cross-walking of scores. (Most of the linking reports based on the PROMIS Wave 1 dataset are included in Volume 1)(Choi et al., 2012).

2.2 The NIH Toolbox for Assessment of Neurological and Behavioral Function (NIH Toolbox)

Developed in 2006 with the NIH Blueprint funding for Neuroscience Research, four domains of assessment central to neurological and behavioral function were created to measure cognition, sensation, motor functioning, and emotional health. The NIH Toolbox for Assessment of Neurological and Behavioral Function(Gershon, 2007) provides investigators with a brief, yet comprehensive measurement tool for assessment of cognitive function, emotional health, sensory and motor function. It provides an innovative approach to measurement that is responsive to the needs of researchers in a variety of settings, with a particular emphasis on measuring outcomes in clinical trials and functional status in large cohort studies, e.g. epidemiological studies and longitudinal studies. Included as subdomains of emotional health were negative affect, psychological well-being, stress and self-efficacy, and social relationships. Three PROMIS emotional distress item banks (Anger, Anxiety, and Depression) were used as measures of negative affect. Additionally, existing “legacy” measures, e.g., Patient Health Questionnaire (PHQ-9) and Center for Epidemiological Studies Depression Scale (CES-D), were flagged as potential candidates for the NIH Toolbox battery because of their history, visibility, and research legacy. Among these legacy measures, we focused on those that were available without proprietary restrictions for research applications. In most cases, these measures had been developed using classical test theory.

2.3 Quality of Life Outcomes in Neurological Disorders (Neuro-QoL)

The National Institute of Neurological Disorders and Stroke sponsored a multi-site project to develop a clinically relevant and psychometrically robust Quality of Life (QOL) assessment tool for adults and children with neurological disorders. The primary goal of this effort, known as Neuro-QoL("Neuro-Qol - Quality of Life Outcomes in Neurological Disorders," 2008), was to enable clinical researchers to compare the QOL impact of different interventions within and across various conditions. This resulted in 13 adult QOL item banks (Anxiety, Depression, Fatigue, Upper Extremity Function - Fine Motor, Lower Extremity Function - Mobility, Applied Cognition - General Concerns, Applied Cognition - Executive Function, Emotional and Behavioral Dyscontrol, Positive Affect and Well-Being, Sleep Disturbance, Ability to Participate in Social Roles and Activities, Satisfaction with Social Roles and Activities, and Stigma).

2.4 PROsetta Stone Data Collection

The National Institutes of Health/National Cancer Institute supported three waves of data collection as part of the PROseta Stone project. The specific aim of each data collection was to administer a range of PROMIS instruments along with legacy measures, following a single sample design (Kolen & Brennan, 2004). For adults (Waves 1 and 2), the assessed (sub)domains comprised negative affect (anger, anxiety, and depression), fatigue, cognitive function, global health, pain interference, physical function, satisfaction with social relationships and activities, sleep disturbance, sleep-related impairment, positive affect and well-being. For children (Wave 3), the following (sub)domains were assessed: anxiety, depression, fatigue, cognitive function, peer relationships, and physical function. The PROsetta Stone data collection allowed investigators to make links to commonly used instruments not administered in PROMIS, Neuro-QoL, and NIH Toolbox studies.

3. Legacy Instruments

Typically, we have linked widely accepted “legacy” measures that were part of the initial validation work for PROMIS or NIH Toolbox. In some cases, instruments were administered as part of the PROsetta Stone project for specific linking purposes. Data were collected on reference measures (e.g., PROMIS Depression) from a minimum of 500 respondents (for stable item parameter estimation), along with responses to at least one other conceptually similar scale or bank to be linked to the reference measure. (See Table 5.1).

3.7 Patient Health Questionnaire-2 (PHQ-2)

The Patient Health Questionnaire-2 (PHQ-2) comprises the first two items of the nine-item PHQ depression module or PHQ-9. The PHQ-2 inquires about the frequency of depressed mood and anhedonia over the past two weeks and is used as a screener rather than to diagnose a depressive disorder or to measure depression severity. Further evaluation with the PHQ-9 is recommended for patients who screen positive in the PHQ-2 assessment. A PHQ-2 score ranges from 0 to 6, with each item scoring as 0 ("not at all") to 3 ("nearly every day"). A score of 3 is considered the optimal cutoff point for screening purposes (Kroenke, Spitzer, & Williams, 2003).

4. Linking Methods

PROMIS full-bank administration allows for single group linking. This linking method is used when two or more measures are administered to the same group of people. For example, two PROMIS banks (Anxiety and Depression) and three legacy measures (MASQ, CES-D, and SF-36/MH) were administered to a sample of 925 people. The order of measures was randomized

so as to minimize potential order effects. The original purpose of the full-bank administration study was to establish initial validity evidence (e.g., validity coefficients), not to establish linking relationships. Some of the measures revealed severely skewed score distributions in the full-bank administration sample and the sample size was relatively small, which might be limiting factors when it comes to determining the linking method. Another potential issue is related to how the non-PROMIS measures are scored and reported. For example, all SF-36 subscales are scored using a proprietary scoring algorithm and reported as normed scores (0 to 100). Others are scored and reported using simple raw summed scores. All PROMIS measures are scored using the final re-centered item response theory (IRT) item parameters and transformed to the T-score metric (mean=50, SD=10).

PROMIS's T-score distributions are standardized such that a score of 50 represents the average (mean) for the US general population, and the standard deviation around that mean is 10 points. A high PROMIS score always represents more of the concept being measured. Thus, for example, a person who has a T-score of 60 is one standard deviation higher than the general population for the concept being measured. For symptoms and other negatively-worded concepts like pain, fatigue, and anxiety, a score of 60 is one standard deviation worse than average; for functional scores and other positively-worded concepts like physical or social function, a score of 60 is one standard deviation better than average, etc.

In order to apply the linking methods consistently across different studies, linking/concordance relationships will be established based on the raw summed score metric of the measures. Furthermore, the direction of linking relationships to be established will be from legacy to PROMIS. That is, each raw summed score on a given legacy instrument will be mapped to a T-score of the corresponding PROMIS instrument/bank. Finally, the raw summed score for each legacy instrument was constructed such that higher scores represent higher levels of the construct being measured. When the measures were scaled in the opposite direction, we reversed the direction of the legacy measure in order for the correlation between the measures to be positive and to facilitate concurrent calibration. As a result, some or all item response scores for some legacy instruments will need to be reverse-coded.

4.1 IRT Linking

One of the objectives of the current linking analysis is to determine whether or not the non-PROMIS measures can be added to their respective PROMIS item bank without significantly altering the underlying trait being measured. The rationale is twofold: (1) the augmented PROMIS item banks might provide more robust coverage both in terms of content and difficulty; and (2) calibrating the non-PROMIS measures on the corresponding PROMIS item bank scale might facilitate subsequent linking analyses. At least, two IRT linking approaches are applicable under the current study design; (1) linking separate calibrations through the Stocking-Lord method and (2) fixed parameter calibration.

Linking separate calibrations might involve the following steps under the current setting.

- First, simultaneously calibrate the combined item set (e.g., PROMIS Depression bank and CES-D).
- Second, estimate linear transformation coefficients (additive and multiplicative constants) using the item parameters for the PROMIS bank items as anchor items.
- Third, transform the metric for the non-PROMIS items to the PROMIS metric.

The second approach, fixed parameter calibration, involves fixing the PROMIS item parameters at their final bank values and calibrating only non-PROMIS items so that the non-PROMIS item parameters may be placed on the same metric as the PROMIS items. The focus is on placing the parameters of non-PROMIS items on the PROMIS scale. Updating the PROMIS item parameters is not desired because the linking exercise is built on the stability of these calibrations. Note that IRT linking would be necessary when the ability level of the full-bank testing sample is different from that of the PROMIS scale-setting sample. If it is assumed that the two samples are from the same population, linking is not necessary and calibration of the items (either separately or simultaneously) will result in item parameter estimates that are on the same scale without any further scale linking. Even though the full-bank testing sample was a subset of the full PROMIS calibration sample, it is still possible that the two samples are somewhat disparate due to some non-random component of the selection process. Moreover, there is some evidence that linking can improve the accuracy of parameter estimation even when linking is not necessary (e.g., two samples are from the same population having the same or similar ability levels). Thus, conducting IRT linking would be worthwhile.

Once the non-PROMIS items are calibrated on the corresponding PROMIS item bank scale, the augmented item bank can be used for standard computation of IRT scaled scores from any subset of the items, including computerized adaptive testing (CAT) and creating short forms. The non-PROMIS items will be treated the same as the existing PROMIS items. Again, the above options are feasible only when the dimensionality of the bank is not altered significantly (i.e., where a unidimensional IRT model is suitable for the aggregate set of items). Thus, prior to conducting IRT linking, it is important to assess dimensionality of the measures based on some selected combinations of PROMIS and non-PROMIS measures. Various dimensionality assessment tools can be used including a confirmatory factor analysis, disattenuated correlations, and essential unidimensionality.

4.2 Equipercentile Linking

The IRT Linking procedures described above are permissible only if the traits being measured are not significantly altered by aggregating items from multiple measures. One potential issue might be creating multidimensionality as a result of aggregating items measuring different traits. For two scales that measure distinct but highly related traits, predicting scores on one scale from those of the other has been used frequently. Concordance tables between PROMIS and non-PROMIS measures can be constructed using equipercentile equating (Kolen & Brennan,

2004; Lord, 1982) when there is insufficient empirical evidence that the instruments measure the same construct. An equipercentile method estimates a nonlinear linking relationship using percentile rank distributions of the two linking measures. The equipercentile linking method can be used in conjunction with a presmoothing method such as the loglinear model (Hanson, Zeng, & Colton, 1994). The frequency distributions are first smoothed using the loglinear model and then equipercentile linking is conducted based on the smoothed frequency distributions of the two measures. Smoothing can also be done at the backend on equipercentile equivalents and is called postsmoothing(Brennan, 2004; Kolen & Brennan, 2004)The cubic-spline smoothing algorithm (Reinsch, 1967) is used in the LEGS program(Brennan, 2004). Smoothing is intended to reduce sampling error involved in the linking process. A successful linking procedure will provide a conversion (crosswalk) table, in which, for example, raw summed scores on the PHQ-9 measure are transformed to the T-score equivalents of the PROMIS Depression measure.

Under the current context, equipercentile crosswalk tables can be generated using two different approaches. First is a direct linking approach where each raw summed score on non-PROMIS measure is mapped directly to a PROMIS T-score. That is, raw summed scores on the non-PROMIS instrument and IRT scaled scores on the PROMIS (reference) instrument are linked directly, although raw summed scores and IRT scaled score have distinct properties (e.g., discrete vs. continuous). This approach might be appropriate when the reference instrument is either an item bank or composed of a large number of items and so various subsets (static or dynamic) are likely to be used but not the full bank in its entirety (e.g., PROMIS Physical Function bank with 124 items). Second is an indirect approach where raw summed scores on the non-PROMIS instrument are mapped to raw summed scores on the PROMIS instrument; and then the resulting raw summed score equivalents are mapped to corresponding scaled scores based on a raw-to-scale score conversion table. Because the raw summed score equivalents may take fractional values, such a conversion table will need to be interpolated using statistical procedures (e.g., cubic spline).

Finally, when samples are small or inadequate for a specific method, random sampling error becomes a major concern (Kolen & Brennan, 2004).. That is, substantially different linking relationships might be obtained if linking is conducted repeatedly over different samples. The type of random sampling error can be measured by the standard error of equating (SEE), which can be operationalized as the standard deviation of equated scores for a given raw summed score over replications(Lord, 1982).

4.3 Linking Assumptions

In Section 5 of this PROsetta Stone report, we present the results of a large number of linking studies using a combination of newly collected and secondary data sets. In most cases, we have applied all three linking methods described in sections 4.1 and 4.2. Our purpose is to provide the maximum amount of useful information. However, the suitability of these methods depends upon the meeting of various linking assumptions. These assumptions require that the

two instruments to be linked measure the same construct, show a high correlation, and are relatively invariant in subpopulation differences (Dorans, 2007). The degree to which these assumptions are met varies across linking studies. Given that different researchers may interpret these requirements differently, we have taken a liberal approach for inclusion of linkages in this book. Nevertheless, we recommend that researchers diagnostically review the classical psychometrics and CFA results in light of these assumptions prior to any application of the cross-walk charts or legacy parameters to their own data. Having investigated a large number of possible links between PROMIS measures and legacy measures, we did apply a few minimal exclusion rules before linking. For example, we generally did not proceed with planned linking when the raw score correlation between two instruments was less than .70.

5. Linking Results

Table 5.1 lists the linking analyses included in this report, which have been conducted based on samples from a NIH Toolbox study (see Section 2 for more details). In most cases, PROMIS instruments were used as the reference (i.e., scores on non-PROMIS instruments are expressed on the PROMIS score metric).

Table 5.1. Linking by Reference Instrument

Section	PROMIS Instrument	Instrument to Link	Study
5.1	PROMIS Depression	PHQ-2	NIH Toolbox CV

5.10 PROMIS Depression and PHQ-2

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression item bank (a selection of 20 highly informative items) and PHQ-2 (2 items). Both instruments were scaled such that higher scores represent higher levels of depression. We excluded 1 participant because of missing responses, leaving a final sample of N=748. We created raw summed scores for each of the measures separately and then for the combined. Summing of item scores assumes that all items have positive correlations with the total as examined in the section on Classical Item Analysis.

5.10.1 Raw Summed Score Distribution

The maximum possible raw summed scores were 100 for PROMIS Depression and 8 for PHQ-2. Figure 5.10.1 and Figure 5.10.2 graphically display the raw summed score distributions of the two measures. Figure 5.10.3 shows the distribution for the combined. Figure 5.10.4 is a scatter plot matrix showing the relationship of each pair of raw summed scores. Pearson correlations are shown above the diagonal. The correlation between PROMIS Depression and PHQ-2 was 0.78. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and PHQ-2 was 0.86. The correlations between the combined score and the measures were 1 and 0.81 for PROMIS Depression and PHQ-2, respectively. Our sample consisted of 748 participants.

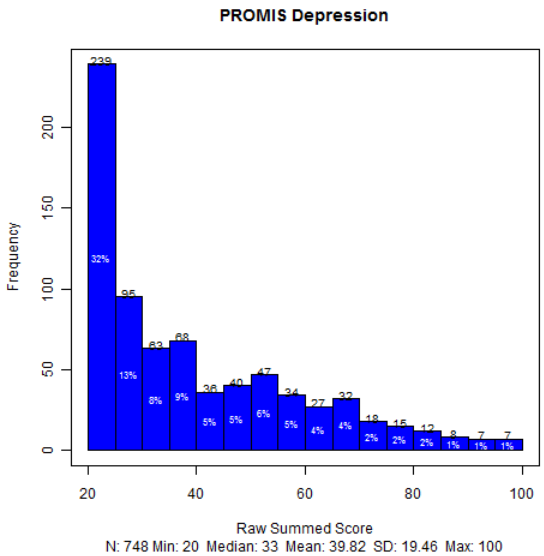

Figure 5.10.1: Raw Summed Score Distribution - PROMIS Depression

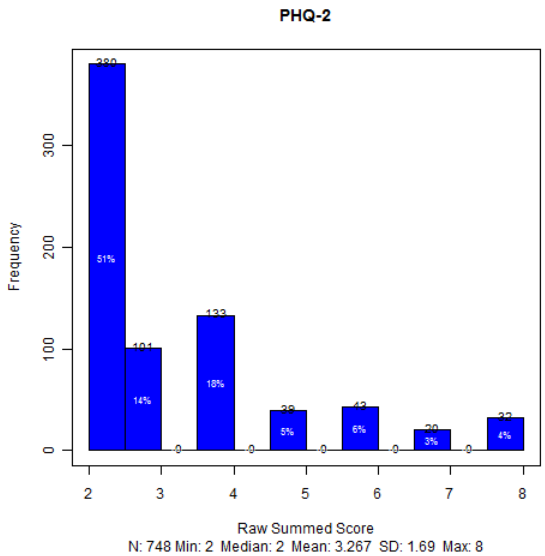

Figure 5.10.2: Raw Summed Score Distribution – PHQ-2

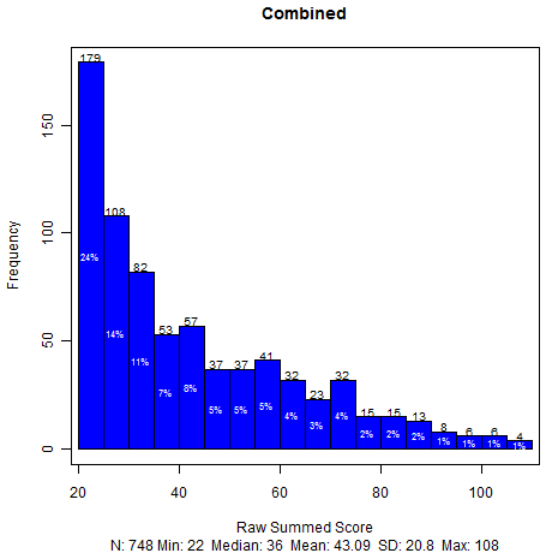

Figure 5.10.3: Raw Summed Score Distribution – Combined

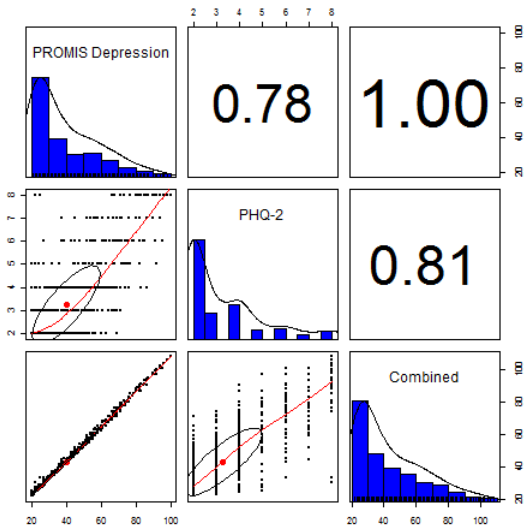

Figure 5.10.4: Scatter Plot Matrix of Raw Summed Scores

5.10.2 Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.10.1 summarizes the results. For PROMIS Depression, Cronbach's alpha internal consistency reliability estimate was 0.979 and adjusted (corrected for overlap) item-total correlations ranged from 0.741 to 0.88. For PHQ-2, alpha was 0.855 and adjusted item-total correlations ranged from 0.747 to 0.747. For the 22 items, alpha was 0.979 and adjusted item-total correlations ranged from 0.682 to 0.881.

Table 5.10.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Depression	20	0.979	0.741	0.826	0.880
PHQ-2	2	0.855	0.747	0.747	0.747
Combined	22	0.979	0.682	0.819	0.881

5.10.3 Confirmatory Factor Analysis (CFA)

To assess the dimensionality of the measures, a categorical confirmatory factor analysis (CFA) was carried out using the WLSMV estimator of Mplus on a subset of cases without missing

responses. A single factor model (based on polychoric correlations) was run on PROMIS Depression and on the combined item set. Table 5.10.2 summarizes the model fit statistics.

Table 5.10.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Depression	20	748	0.988	0.986	0.089
Combined	22	748	0.984	0.983	0.093

5.10.4 Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 22 items according to the graded response model. The calibration was run using `MULTILOG` and two different approaches as described previously (i.e., IRT linking vs. fixed-parameter calibration). For IRT linking, all 22 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 20 PROMIS Depres- sion items served as anchor items to transform the item parameter estimates for the PHQ-2 items onto the PROMIS Depression metric. We used four IRT linking methods implemented in `plink` (Weeks, 2010): mean/mean, mean/sigma, Haebara, and Stocking-Lord. The first two methods are based on the mean and standard deviation of item parameter estimates, whereas the latter two are based on the item and test information curves. Table 5.10.3 shows the additive (A) and multiplicative (B) transformation constants derived from the four linking methods. For fixed-parameter calibration, the item parameters for the PROMIS Depression items were constrained to their final bank values, while the PHQ-2 items were calibrated, under the constraints imposed by the anchor items.

Table 5.10.3: IRT Linking Constants

	A	B
Mean/Mean	1.156	0.343
Mean/Sigma	1.217	0.298
Haebara	1.220	0.324
Stocking-Lord	1.207	0.310

The item parameter estimates for the PHQ-2 items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.10.3. The PHQ-2 item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Depression metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for PHQ-2 as shown in Figure 5.10.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.10.6 displays the differences on the vertical axis.

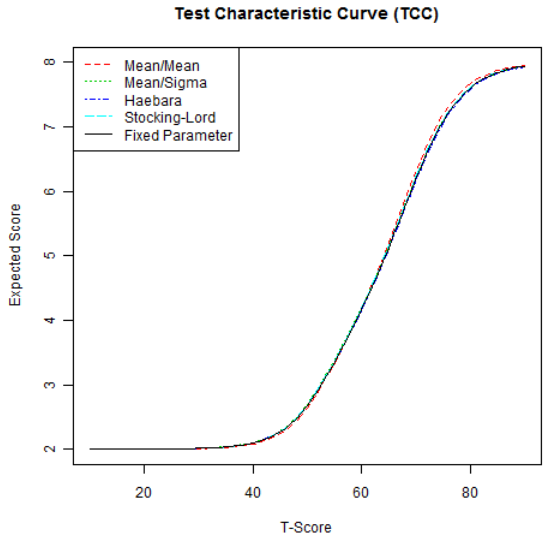

Figure 5.10.5: Test Characteristic Curves (TCC) from Different Linking Methods

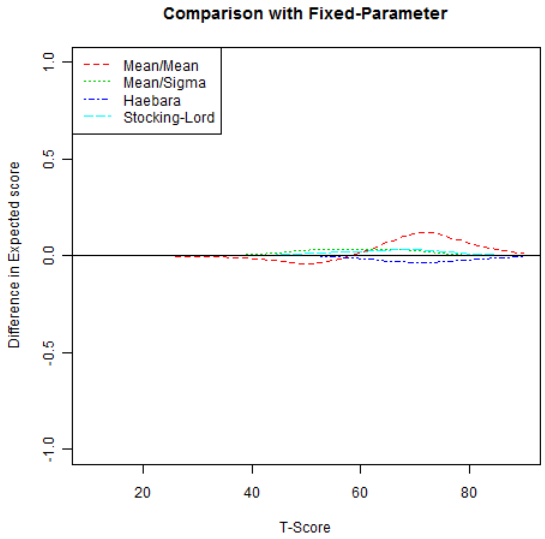

Figure 5.10.6: Difference in Test Characteristic Curves (TCC)

Table 5.10.4 shows the fixed-parameter calibration item parameter estimates for PHQ-2. The marginal reliability estimate for PHQ-2 based on the item parameter estimates was 0.572. The marginal reliability estimates for PROMIS Depression and the combined set were 0.929 and 0.931, respectively. The slope parameter estimates for PHQ-2 ranged from 1.86 to 2.75 with a mean of 2.3. The slope parameter estimates for PROMIS Depression ranged from 2.36 to 4.45 with a mean of 3.26. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.10.7 displays the scale information functions for PROMIS Depression, PHQ-2, and the combined set of 22. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.10.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.10.4: Fixed-Parameter Calibration Item Parameter Estimates for PHQ-2

a	cb1	cb2	cb3	NCAT
1.862	0.471	1.689	2.305	4
2.748	0.310	1.443	2.120	4

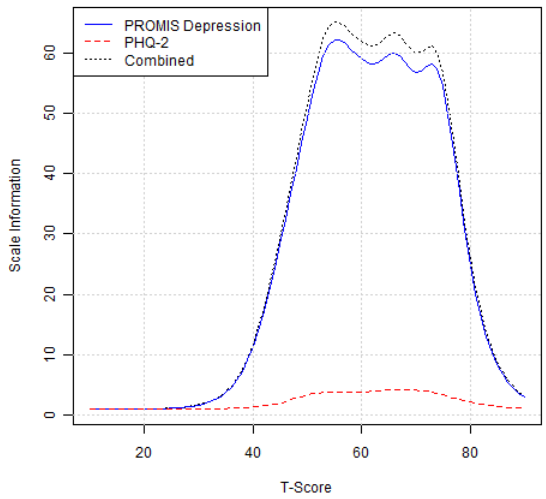

Figure 5.10.7: Comparison of Scale Information Functions

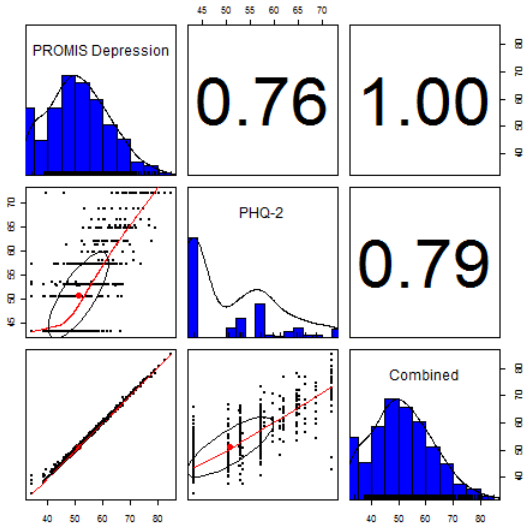

Figure 5.10.8: Comparison of IRT Scaled Scores

5.10.5 Raw Score to T-Score Conversion using Linked IRT Parameters

The IRT model implemented in PROMIS (i.e., the graded response model) uses the pattern of item responses for scoring, not just the sum of individual item scores. However, a crosswalk table mapping each raw summed score point on PHQ-2 to a scaled score on PROMIS Depression can be useful. Based on the PHQ-2 item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 28 can be used to map simple raw summed scores from PHQ-2 to T-score values linked to the PROMIS Depression metric. Each raw summed score point and corresponding PROMIS scaled score are presented along with the standard error associated with the scaled score. The raw summed score is constructed such that for each item, consecutive integers in base 1 are assigned to the ordered response categories.

5.10.6 Equipercentile Linking

We mapped each raw summed score point on PHQ-2 to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on PHQ-2. Theoretically, the equipercentile linking function is symmetrical for continuous random variables (X and Y). Therefore, the linking function for the values in X to those in Y is the same as that for the values in Y to those in X. However, for discrete variables like raw summed scores the equipercentile linking functions can be slightly different (due to rounding errors and differences in score ranges) and hence may need to be

obtained separately. Figure 5.10.9 displays the cumulative distribution functions of the measures. Figure 5.10.10 shows the equipercentile linking functions based on raw summed scores, from PHQ-2 to PROMIS Depression. When the number of raw summed score points differs substantially, the equipercentile linking functions could deviate from each other noticeably. The problem can be exacerbated when the sample size is small. Appendix Table 29 and Appendix Table 30 show the equipercentile crosswalk tables. The result shown in Appendix Table 29 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 30 shows the result based on the indirect (raw summed score to raw summed score equivalent to scaled score equivalent) approach (Refer to Section 4.2 for details). Three separate equipercentile equivalents are presented: one is equipercentile without post smoothing ("Equipercentile Scale Score Equivalents") and two with different levels of postsmoothing, i.e., "Equipercentile Equivalents with Postsmoothing (Less Smoothing)" and "Equipercentile Equivalents with Postsmoothing (More Smoothing)." Postsmoothing values of 0.3 and 1.0 were used for "Less" and "More," respectively (Refer to Brennan, 2004 for details).

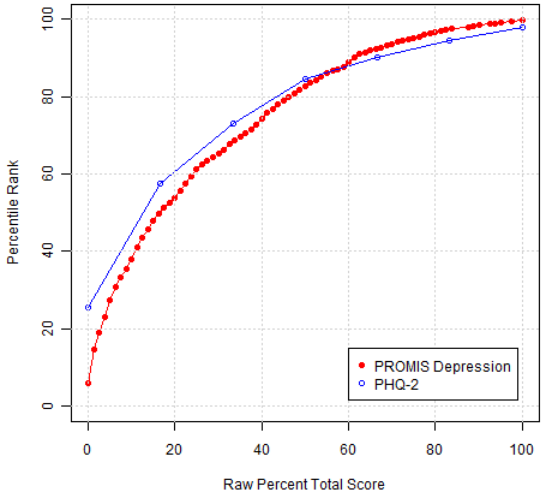

Figure 5.10.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

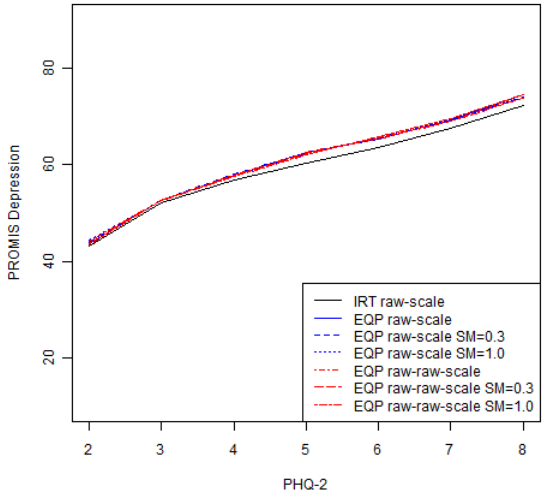

Figure 5.10.10: Equipercentile Linking Functions

5.10.7 Summary and Discussion

The purpose of linking is to establish the relationship between scores on two measures of closely related traits. The relationship can vary across linking methods and samples employed. In equipercentile linking, the relationship is determined based on the distributions of scores in a given sample. Although IRT-based linking can potentially offer sample-invariant results, they are based on estimates of item parameters, and hence subject to sampling errors. A potential issue with IRT-based linking methods is, however, the violation of model assumptions as a result of

combining items from two measures (e.g., unidimensionality and local independence). As displayed in Figure 5.10.10, the relationships derived from various linking methods are consistent, which suggests that a robust linking relationship can be determined based on the given sample.

To further facilitate the comparison of the linking methods, Table 5.10.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Depression T-scores and PHQ-2 scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.10.10), the method labeled "IRT pattern scoring" refers to IRT scoring based on the pattern of item responses instead of raw summed scores. With respect to the correlation between observed and linked T-scores, IRT pattern scoring produced the best result (0.763), followed by EQP raw-raw-scale SM=1.0 (0.748). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (7.135), followed by IRT raw-scale (7.325).

Table 5.10.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.763	0.276	7.135	7.135
IRT raw-scale	0.748	0.381	7.320	7.325
EQP raw-scale SM=0.0	0.748	-0.424	7.389	7.396
EQP raw-scale SM=0.3	0.748	-0.637	7.357	7.379
EQP raw-scale SM=1.0	0.748	-0.883	7.332	7.381
EQP raw-raw-scale SM=0.0	0.748	-0.333	7.402	7.404
EQP raw-raw-scale SM=0.3	0.748	-0.438	7.368	7.376
EQP raw-raw-scale SM=1.0	0.748	-0.649	7.352	7.376

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=748) over a large number of replications (i.e., 10,000).

Table 5.10.6 summarizes the mean and standard deviation of differences between the observed and linked T-scores by linking method and sample size. For each replication, the mean difference between the observed and equated PROMIS Depression T-scores was computed. Then the mean and the standard deviation of the means were computed over replications as bias and empirical standard error, respectively. As the sample size increased (from 25 to 75), the empirical standard error decreased steadily. At a sample size of 75, IRT pattern scoring produced the smallest standard error, 0.781. That is, the difference between the mean PROMIS Depression T-score and the mean equated PHQ-2 T-score based on a similar sample of 75 cases is expected to be around ±1.56 (i.e., 2 × 0.781).

Table 5.10.6: Comparison of Resampling Results

Methods	Mean (N=25)	SD (N=25)	Mean (N=50)	SD (N=50)	Mean (N=75)	SD (N=75)
IRT pattern scoring	0.256	1.402	0.275	0.976	0.287	0.781
IRT raw-scale	0.393	1.414	0.374	1.001	0.397	0.793
EQP raw-scale SM=0.0	-0.433	1.462	-0.413	1.007	-0.425	0.809
EQP raw-scale SM=0.3	-0.623	1.453	-0.650	1.016	-0.631	0.804
EQP raw-scale SM=1.0	-0.884	1.453	-0.887	0.989	-0.889	0.799
EQP raw-raw-scale SM=0.0	-0.332	1.448	-0.338	1.015	-0.351	0.811
EQP raw-raw-scale SM=0.3	-0.438	1.434	-0.435	1.017	-0.442	0.809
EQP raw-raw-scale SM=1.0	-0.648	1.449	-0.658	1.009	-0.640	0.809

Examining a number of linking studies in the current project revealed that the two linking methods (IRT and equipercentile) in general produced highly comparable results. Some noticeable discrepancies were observed (albeit rarely) in some extreme score levels where data were sparse. Model-based approaches can provide more robust results than those relying solely on data when data is sparse. The caveat is that the model should fit the data reasonably well. One of the potential advantages of IRT-based linking is that the item parameters on the linking instrument can be expressed on the metric of the reference instrument, and therefore can be combined without significantly altering the underlying trait being measured. As a result, a larger item pool might be available for computerized adaptive testing or various subsets of items can be used in static short forms. Therefore, IRT-based linking (Appendix Table 28) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

6. Appendix Table 28: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PHQ-2 to PROMIS Depression (NIH Toolbox Study) - RECOMMENDED

PHQ-2 Score	PROMIS Depression T-score	SE
0	43.1	7.2
1	52.0	5.7
2	56.9	5.3
3	60.2	5.9
4	63.5	5.6
5	67.5	5.2
6	72.2	5.8

7. Appendix Table 29: Direct (Raw to Scale) Equipercentile Crosswalk Table – From PHQ-2 to PROMIS Depression – Note: Table 28 is recommended.

PHQ-2 Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	44	44	44	0.87
1	53	53	53	0.39
2	58	58	58	0.76
3	63	62	62	0.80
4	65	65	66	0.44
5	69	69	69	1.87
6	74	74	74	2.57

8. Appendix table 30: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From PHQ-2 to PROMIS Depression – Note: Table 28 is recommended.

PHQ-2 Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	43	44	44
1	53	53	53
2	58	58	58
3	62	62	62
4	65	66	66
5	69	69	70
6	74	74	74