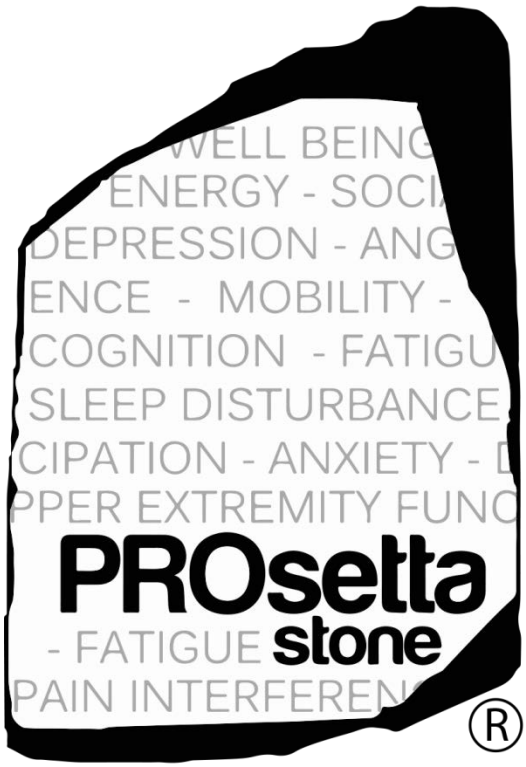

PROSETTA STONE® ANALYSIS REPORT

A ROSETTA STONE FOR PATIENT REPORTED OUTCOMES

SEUNG W. CHOI, TRACY PODRABSKY, NATALIE MCKINNEY, BENJAMIN D. SCHALET, KARON F. COOK & DAVID CELLA

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 preliminary report encompasses the first wave of 20 linking studies based on available PRO data from PROMIS (aka, PROMIS Wave I), Toolbox, and Neuro-QOL.

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's (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 Therapy-Fatigue (FACIT-F), Brief Pain Inventory (BPI), and SF-36. In addition to the pairs for validity (e.g., PROMIS Depression and CES-D), the PROMIS Wave I data allows for the potential for linking over a dozen pairs of measures/subscales. Furthermore, included within each of the PROMIS banks were items from many 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.

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

2.2. The NIH Toolbox for Assessment of Neurological and Behavioral Function (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[3] 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 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[3], 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).

3. Legacy Instruments

The following instruments are widely accepted "legacy" measures that have been used as part of the initial validation work for PROMIS and Toolbox. Data were collected on a minimum of 500

[3] www.nihtoolbox.org

respondents (for stable item parameter estimation) along with at least one other conceptually similar scale or bank.

3.1. Mood and Anxiety Symptom Questionnaire (MASQ)

The Mood and Anxiety Symptom Questionnaire (MASQ) is a 77-item self-report questionnaire that assesses depressive, anxious, and mixed symptomatology. Three scales measure General Distress: depressive symptoms (12 items), anxious symptoms (11 items), and mixed symptoms (15 items). There are also anxiety-specific (Anxious Arousal, 17 items) and depression-specific scales (Anhedonic Depression, 22 items). Higher scores reflect greater levels of symptomatology. (Watson et al., 1995). For the current analysis, we used the Anxious Symptoms scale.

3.2. SF-36

The SF-36 is a multi-purpose, short-form health survey with 36 items. It yields an 8-scale profile of functional health and well-being scores as well as psychometrically-based physical and mental health summary scores and a preference-based health utility index. The SF-36 version 2 (Ware, Kosinski, & Dewey, 2000.) consists of items assessing physical functioning (PF; 10 items), social functioning (SF; 2 items), role limitation due to physical health (RP; 4 items), bodily pain (BP; 2 items), mental health (MH; 5 items), role limitations due to emotional health (RE; 3 items), vitality (VT; 4 items), general health perceptions (GH; 5 items), and reported health transition (1 item). The Physical Component Score (PCS) and Mental Component Score (MCS) range from 0-100 with higher scores indicating better health-related quality of life.

3.3. Center for Epidemiological Studies Depression Scale (CES-D)

The Center for Epidemiological Studies Depression Scale (CES-D) is a 20-item measure designed to assess depressive symptoms in the general population. Items are rated for the past week using a four-point scale for duration (from “rarely or none of the time” to “most or all of the time”). The CES-D has good psychometric properties and has been used in a variety of contexts, including community samples and clinical samples with both medical and psychiatric illnesses (Radloff, 1977).

3.4. Buss-Perry Aggression Questionnaire (BPAQ)

The Buss-Perry Aggression Questionnaire (BPAQ) is a 29-item self-report measure that includes four subscales: physical aggression (9 items), verbal aggression (5 items), anger (7 items), and hostility (8 items) (Buss & Perry, 1992). There is no time frame specified, and items are rated using a seven-point scale from “extremely uncharacteristic” to “extremely characteristic”.

3.5. Health Assessment Questionnaire (HAQ)

The Health Assessment Questionnaire (HAQ) was developed as a comprehensive measure of outcomes in patients with a wide variety of rheumatic diseases (Fries, Spitz, Kraines, & Holman, 1980). It should be considered a generic rather than a disease-specific instrument. The HAQ has been administered primarily in one of two versions, short HAQ-DI (Disability Index) or the Full HAQ. The HAQ-DI assesses the extent of a patient’s functional ability. It is composed of 20

items in 8 categories (Dressing and Grooming, Hygiene, Arising, Reach, Eating, Grip, Walking, Common Daily Activities).

3.6. Functional Assessment of Chronic Illness Therapy (FACIT)

The Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System is a collection of QOL questionnaires targeted to the management of chronic illness including cancer. The FACT-G (now in Version 4) is a 27-item compilation of general questions divided into four subscales: Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, and Functional Well-Being. It is considered appropriate for use with patients with any form of cancer, and has also been used and validated in other chronic illness conditions (e.g., HIV/AIDS, multiple sclerosis) and in the general population (using a slightly modified version). Validation of a core measure allowed for the evolution of multiple disease, treatment, condition, and non-cancer-specific subscales. FACIT subscales are constructed to complement the FACT-G, addressing relevant disease-, treatment-, or condition-related issues not already covered in the general questionnaire. Each is intended to be as specific as necessary to capture the clinically-relevant problems associated with a given condition or symptom, yet general enough to allow for comparison across diseases, and extension, as appropriate, to other chronic medical conditions. For the current analysis, we used the Fatigue scale. The Functional Assessment of Chronic Illness Therapy-Fatigue Scale (FACIT-Fatigue scale) is a 13-item questionnaire that assesses self-reported fatigue and its impact upon daily activities and function (Yellen, Cella, Webster, Blendowsky, & Kaplan, 1997). It was developed to meet a growing demand for the precise evaluation of fatigue associated with anemia in cancer patients. Subsequently, it has been employed in over 70 published studies including over 20,000 people, including cancer patients receiving chemotherapy (Berndt et al., 2005; Quirt et al., 2001), cancer patients not receiving chemotherapy (Quirt et al., 2001; Quirt et al., 2002), long term cancer survivors (Ng et al., 2005), childhood cancer survivors (Mulrooney et al., 2008), rheumatoid arthritis (Cella et al., 2005; Mease et al., 2008; Mittendorf et al., 2007), psoriatic arthritis (Chandran, Bhella, Schentag & Gladman, 2007), paroxysmal nocturnal hemoglobinuria (Brodsky et al., 2008), and Parkinson's disease (Hagell et al., 2006). It has also been validated in the general United States population (Brucker, Yost, Cashy, Webster & Cella., 2005; Cella, Lai, Chang, Peterman & Slavin, 2002). In all cases, the FACIT-Fatigue scale has been found to be reliable and valid.

3.7. BPI Severity and Interference

The Brief Pain Inventory (BPI) (Cleeland & Ryan, 1994) produces pain severity and pain interference scores ranging from 0 to 10 and higher scores indicate worse pain. There is a short and a long form. There are 15 questions on the Short Form BPI (9 questions, with the last question containing 7 parts). In PROMIS calibration testing, 11 of the 15 questions were administered (BPI items 1, 2, 7, and 8 were omitted). However, for some BPI items, PROMIS calibration testing used a one week recall period. This matches the recall period used by the BPI long form, but not the 24-hour recall period used in the BPI short form.

3.8. Generalized Anxiety Disorder Scale (GAD-7)

The Generalized Anxiety Disorder Scale (GAD-7) is a 7-item instrument developed with primary care patients and the goal of identifying probable cases of GAD (Spitzer, Kroenke, Williams, &

Löwe, 2006). Items are rated for the last two weeks, using a four-point scale for duration (from "not at all" to "nearly every day").

3.9. Kessler 6 Mental Health Scale (K6)

The Kessler 6 Mental Health Scale (K6) (Kessler et. al., 2003) is a measure of non-specific psychological distress. The K6 is a tool used for screening mental health issues in a general adult population. The scale was designed to be sensitive around the threshold for the clinically significant range of the distribution of non-specific distress in an effort to maximize the ability to discriminate cases of serious mental illness from the rest.

3.10. Patient Health Questionnaire (PHQ-9)

The Patient Health Questionnaire (PHQ-9) is a nine-item instrument designed for use in primary care settings (Kroenke, Spitzer & Williams., 2001). It is based directly on the diagnostic criteria for major depressive disorder in the Diagnostic and Statistical Manual, Fourth Edition (American Psychiatric Association, 2000). Items are rated for the last two weeks, using a four-point scale for duration (from "not at all" to "nearly every day"). The PHQ-9 has been adopted widely as a screening and diagnostic tool as well as a measure for monitoring treatment.

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 (Lord, 1982; Kolen & Brennan, 2004) 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, we present the results of a large number of linking studies using secondary data sets. In each case, 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.

5. Linking Results

Table 5.1 lists the linking analyses included in this report, which have been conducted based on samples from two different studies: PROMIS and Toolbox (see Section 2 for more details). In all cases, PROMIS instruments were used as the reference (i.e., scores on non-PROMIS instruments are expressed on the PROMIS score metric); however, shorter versions of PROMIS were used in Toolbox.

Table 5.1. Linking by Study

Section	Study	PROMIS Instrument	Non-PROMIS Instrument to Link
5.1	PROMIS Wave1	Anxiety	Mood and Anxiety Symptom Questionnaire (MASQ)
5.2	PROMIS Wave1	Anxiety	SF-36 Mental Health (SF-36/MH)
5.3	PROMIS Wave1	Depression	Center for Epidemiological Studies Depression Scale (CES-D)
5.4	PROMIS Wave1	Depression	SF-36 Mental Health (SF-36/MH)
5.5	PROMIS Wave1	Anger	Buss Perry Aggression Questionnaire (BPAQ)
5.6	PROMIS Wave1	Physical Function	Health Assessment Questionnaire (HAQ-DI)
5.7	PROMIS Wave1	Physical Function	SF-36 Physical Functioning (SF-36/PF)
5.8	PROMIS Wave1	Fatigue	Functional Assessment of Chronic Illness Therapy – Fatigue Scale (FACIT-F)
5.9	PROMIS Wave1	Fatigue	SF-36 Vitality (SF-36/VT)
5.10	PROMIS Wave1	Pain Interference	Brief Pain Inventory Severity (BPI Severity)
5.11	PROMIS Wave1	Pain Interference	Brief Pain Inventory Interference (BPI Interference)
5.12	Toolbox	Anxiety	Generalized Anxiety Disorder Scale (GAD-7)
5.13	Toolbox	Anxiety	Kessler 6 Mental Health Scale (K6)
5.14	Toolbox	Anxiety	Mood and Anxiety Symptom Questionnaire (MASQ)
5.15	Toolbox	Depression	Center for Epidemiological Studies Depression Scale (CES-D)
5.16	Toolbox	Depression	Patient Health Questionnaire (PHQ-9)
5.17	Neuro-QOL	Anxiety	Neuro-QOL Anxiety
5.18	Neuro-QOL	Depression	Neuro-QOL Depression
5.19	Neuro-QOL	Physical Function	Neuro-QOL Mobility
5.20	Neuro-QOL	Physical Function	Neuro-QOL Upper Extremity

5.1. PROMIS Anxiety and MASQ

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Anxiety, namely the PROMIS Anxiety item bank (29 items) and MASQ (11 items). PROMIS Anxiety was scaled such that higher scores represent higher levels of Anxiety. 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.

5.1.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 145 for PROMIS Anxiety and 55 for MASQ. Figure 5.1.1 and Figure 5.1.2 graphically display the raw summed score distributions of the two measures. Figure 5.1.3 shows the distribution for the combined. Figure 5.1.4 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 Anxiety and MASQ was 0.85. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anxiety and MASQ was 0.91. The correlations between the combined score and the measures were 0.99 and 0.91 for PROMIS Anxiety and MASQ, respectively.

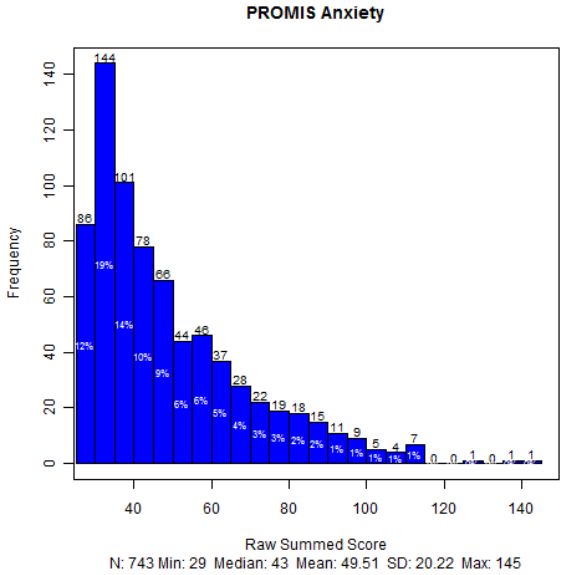

Figure 5.1.1: Raw Summed Score Distribution - PROMIS Instrument

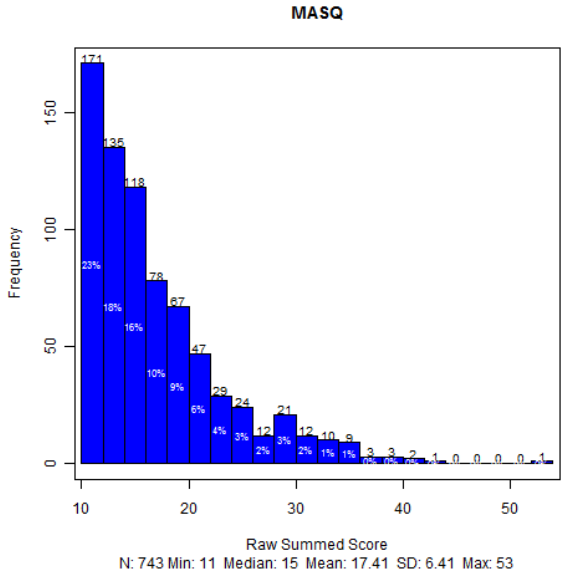

Figure 5.1.2: Raw Summed Score Distribution – Linking Instrument

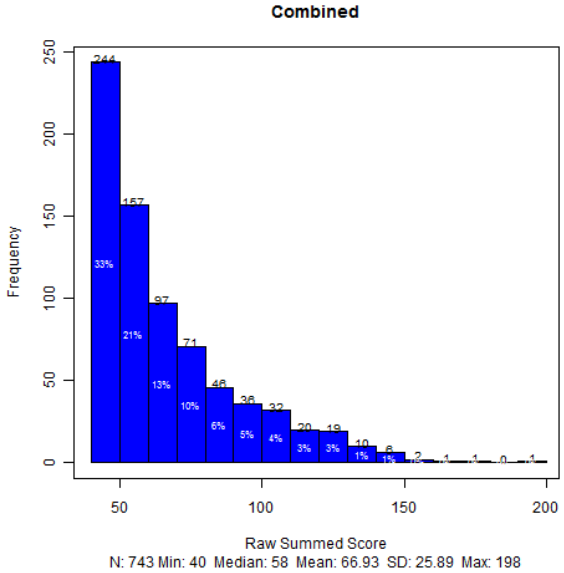

Figure 5.1.3: Raw Summed Score Distribution – Combined

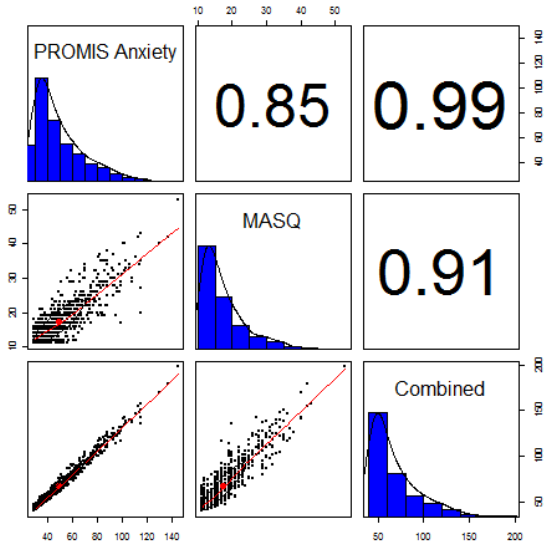

Figure 5.1.4: Scatter Plot Matrix of Raw Summed Scores

5.1.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.1.1 summarizes the results. For PROMIS Anxiety, Cronbach's alpha internal consistency reliability estimate was 0.971 and adjusted (corrected for overlap) item-total correlations ranged from 0.511 to 0.827. For MASQ, alpha was 0.893 and adjusted item-total correlations ranged from 0.433 to 0.779. For the 40 items, alpha was 0.975 and adjusted item-total correlations ranged from 0.391 to 0.832.

Table 5.1.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Anxiety	29	0.971	0.511	0.729	0.827
MASQ	11	0.893	0.433	0.622	0.779
Combined	40	0.975	0.391	0.695	0.832

5.1.3. 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 each of the two measures separately and on the combined. Table 5.1.2 summarizes the model fit statistics. For PROMIS Anxiety, the fit statistics were as follows: CFI = 0.983, TLI = 0.982, and RMSEA = 0.054. For MASQ, CFI = 0.937, TLI = 0.922, and RMSEA = 0.163. For the 40 items, CFI = 0.951, TLI = 0.948, and RMSEA = 0.074. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.1.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Anxiety	29	751	0.983	0.982	0.054
MASQ	11	751	0.937	0.922	0.163
Combined	40	751	0.951	0.948	0.074

5.1.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 40 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 40 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 29 PROMIS Anxiety items served as anchor items to transform the item parameter estimates for the MASQ items onto the PROMIS Anxiety 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 5.1.3 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 items were constrained to their final bank values, while the MASQ items were calibrated under the constraints imposed by the anchor items.

Table 5.1.3: IRT Linking Constants

	A	B
Mean/Mean	1.131	0.290
Mean/Sigma	1.170	0.251
Haebara	1.168	0.248
Stocking-Lord	1.163	0.258

The item parameter estimates for the MASQ items were linked to the PROMIS Anxiety metric using the transformation constants shown in Table 5.1.3. The MASQ item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Anxiety metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for MASQ as shown in Figure 5.1.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.1.6 displays the differences on the vertical axis.

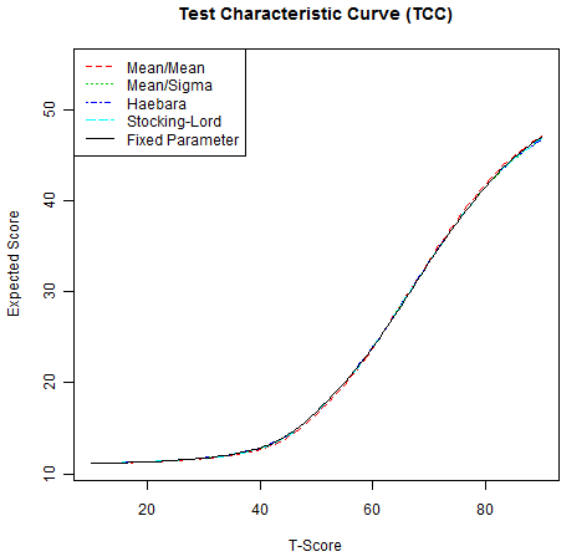

Figure 5.1.5: Test Characteristic Curves (TCC) from Different Linking Methods

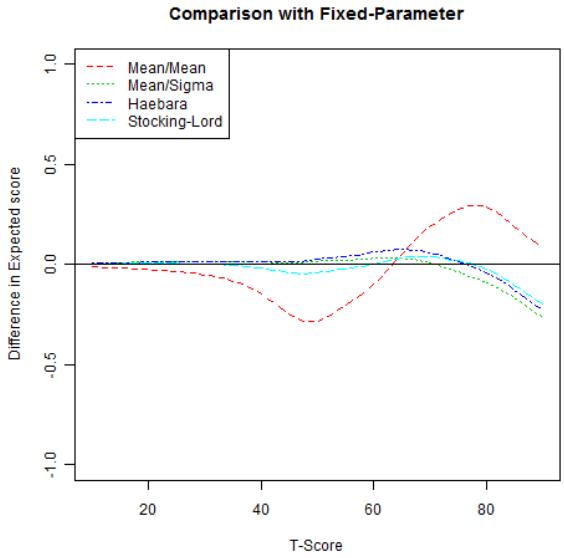

Figure 5.1.6: Difference in Test Characteristic Curves (TCC)

Table 5.1.4 shows the fixed-parameter calibration item parameter estimates for MASQ. The marginal reliability estimate for MASQ based on the item parameter estimates was 0.847. The marginal reliability estimates for PROMIS Anxiety and the combined set were 0.946 and 0.956, respectively. The slope parameter estimates for MASQ ranged from 0.785 to 3.06 with a mean of 1.86. The slope parameter estimates for PROMIS Anxiety ranged from 1.27 to 3.88 with a mean of 2.72. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.1.7 displays the scale information functions for PROMIS Anxiety, MASQ, and the combined set of 40. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.1.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.1.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
2.473	0.629	1.758	2.546	4.161
0.785	1.216	3.100	4.632	7.856
2.979	-0.042	1.192	1.873	3.030
3.063	-0.116	1.053	1.769	2.634
1.253	1.610	3.222	3.786	4.746
1.092	0.246	1.996	3.185	5.490
2.307	-0.117	1.015	1.747	2.745
2.180	-0.014	1.072	1.772	2.782
1.163	0.979	2.508	3.691	5.234
2.343	0.030	1.064	1.640	2.635
0.838	-1.105	1.004	2.743	4.832

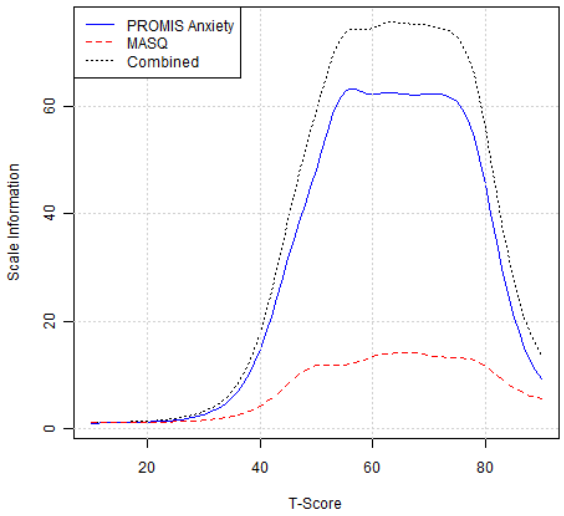

Figure 5.1.7: Comparison of Scale Information Functions

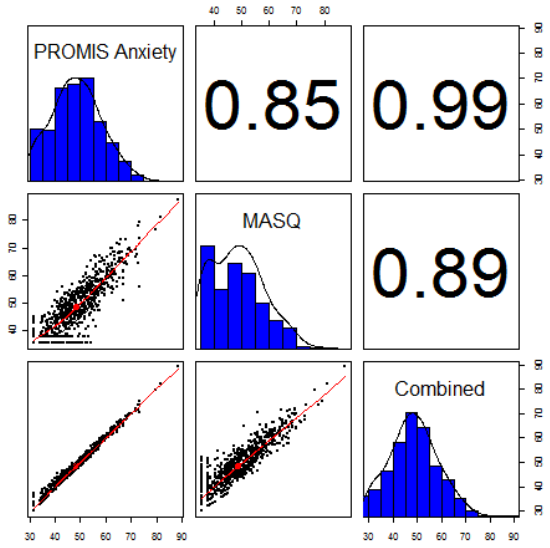

Figure 5.1.8: Comparison of IRT Scaled Scores

5.1.5. 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 MASQ to a scaled score on PROMIS Anxiety can be useful. Based on the MASQ item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 1 can be used to map simple raw summed scores from MASQ to T-score values linked to the PROMIS Anxiety 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.

5.1.6. Equipercentile Linking

We mapped each raw summed score point on MASQ to a corresponding scaled score on PROMIS Anxiety by identifying scores on PROMIS Anxiety that have the same percentile ranks as scores on MASQ. 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 5.1.9 displays the cumulative distribution functions of the measures. Figure 5.1.10 shows the equipercentile linking functions based on raw summed scores, from MASQ to PROMIS Anxiety. 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 2 and Appendix Table 3 show the equipercentile crosswalk tables. The result shown in Appendix Table 2 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 3 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).

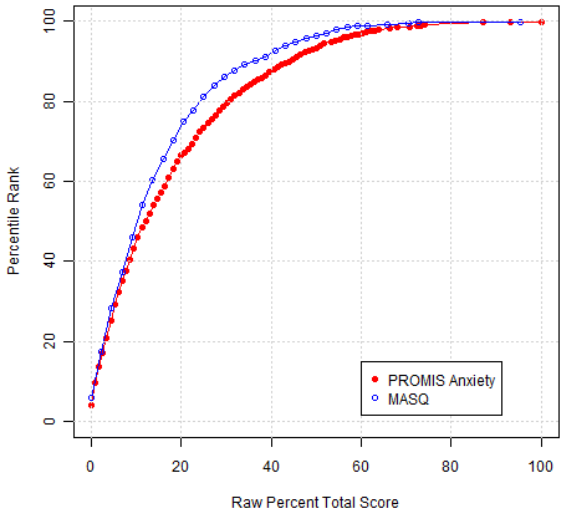

Figure 5.1.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

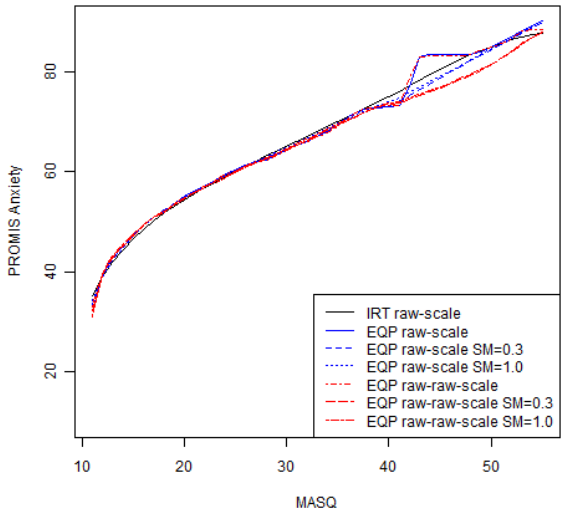

Figure 5.1.10: Equipercentile Linking Functions

5.1.7. 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 5.1.10, 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 5.1.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Anxiety T-scores and MASQ scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.1.10), 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 (0.85), followed by IRT raw-scale (0.818). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.355), followed by IRT raw-scale (5.851).

Table 5.1.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.850	-0.067	5.358	5.355
IRT raw-scale	0.818	0.023	5.855	5.851
EQP raw-scale SM=0.0	0.807	-0.005	6.138	6.134
EQP raw-scale SM=0.3	0.810	-0.053	6.047	6.043
EQP raw-scale SM=1.0	0.815	-0.146	5.956	5.954
EQP raw-raw-scale SM=0.0	0.809	-0.102	6.054	6.051
EQP raw-raw-scale SM=0.3	0.807	-0.013	6.114	6.110
EQP raw-raw-scale SM=1.0	0.804	0.034	6.191	6.187

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=743), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=743) over a large number of replications (i.e., 10,000).

Table 5.1.6 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 Anxiety 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, 0.584. That is, the difference between the mean PROMIS Anxiety T-score and the mean equated MASQ T-score based on a similar sample of 75 cases is expected to be around ±1.17 (i.e., 2 x 0.584).

Table 5.1.6: 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	-0.063	1.063	-0.067	0.740	-0.067	0.584
IRT raw-scale	0.022	1.141	0.017	0.796	0.018	0.636
EQP raw-scale SM=0.0	-0.016	1.200	-0.006	0.828	-0.007	0.673
EQP raw-scale SM=0.3	-0.052	1.176	-0.061	0.823	-0.048	0.661
EQP raw-scale SM=1.0	-0.142	1.174	-0.151	0.818	-0.149	0.647
EQP raw-raw-scale SM=0.0	-0.113	1.198	-0.106	0.835	-0.103	0.667
EQP raw-raw-scale SM=0.3	-0.019	1.202	0.003	0.832	-0.013	0.665
EQP raw-raw-scale SM=1.0	0.029	1.203	0.031	0.840	0.038	0.678

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 1) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.2. PROMIS Anxiety and SF-36/MH

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Anxiety, namely the PROMIS Anxiety (29 items) and SF-36/MH (5 items). PROMIS Anxiety was scaled such that higher scores represent higher levels of anxiety; for the SF-36/MH, higher scores represent lower levels of anxiety. 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.

5.2.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 145 for PROMIS Anxiety and 25 for SF-36/MH. Figure 5.2.1 and Figure 5.2.2 graphically display the raw summed score distributions of the two measures. Figure 5.2.3 shows the distribution for the combined. Figure 5.2.4 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 Anxiety and SF-36/MH was -0.81. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anxiety and SF-36/MH was -0.87. The correlations between the combined score and the measures were 0.99 and 0.87 for PROMIS Anxiety and SF-36/MH, respectively.

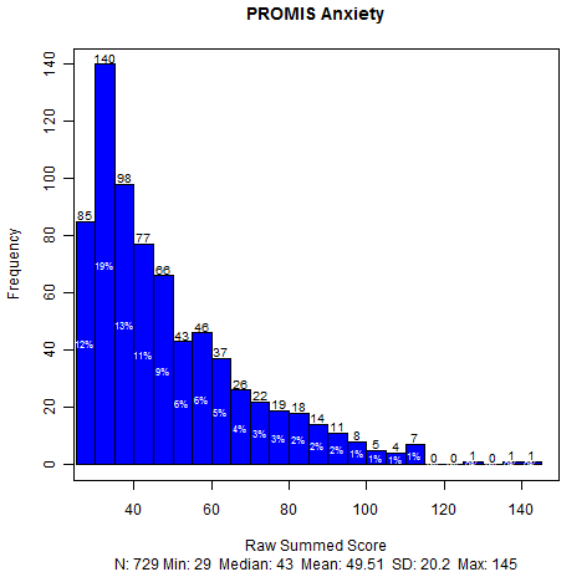

Figure 5.2.1: Raw Summed Score Distribution - PROMIS Instrument

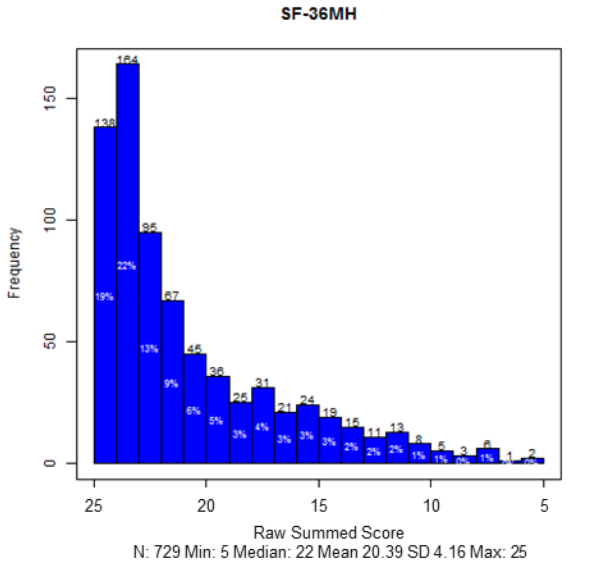

Figure 5.2.2: Raw Summed Score Distribution – Linking Instrument

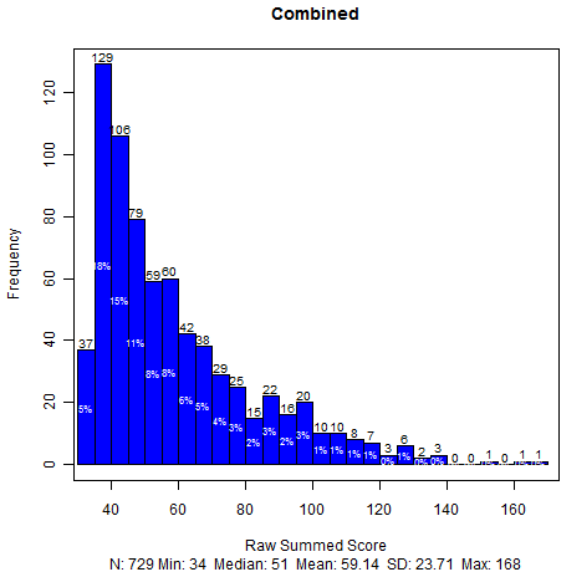

Figure 5.2.3: Raw Summed Score Distribution – Combined

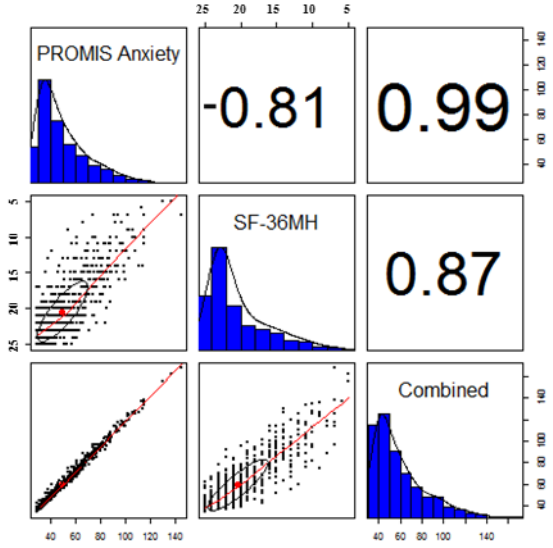

Figure 5.2.4: Scatter Plot Matrix of Raw Summed Scores

5.2.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.2.1 summarizes the results. For PROMIS Anxiety, Cronbach's alpha internal consistency reliability estimate was 0.971 and adjusted (corrected for overlap) item-total correlations ranged from 0.512 to 0.828. For SF-36/MH, alpha was 0.888 and adjusted item-total correlations ranged from 0.578 to 0.818. For the 34 items, alpha was 0.974 and adjusted item-total correlations ranged from 0.504 to 0.833.

Table 5.2.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Anxiety	29	0.971	0.512	0.728	0.828
SF-36/MH	5	0.888	0.578	0.731	0.818
Combined	34	0.974	0.504	0.723	0.833

5.2.3. 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 each of the two measures separately and on the combined. Table 5.2.2 summarizes the model fit statistics. For PROMIS Anxiety, the fit statistics were as follows: CFI = 0.983, TLI = 0.981, and RMSEA = 0.054. For SF-36/MH, CFI = 0.988, TLI = 0.976, and RMSEA = 0.207. For the 34 items, CFI =

0.956, TLI = 0.953, and RMSEA = 0.078. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.2.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Anxiety	29	737	0.983	0.981	0.054
SF-36/MH	5	737	0.988	0.976	0.207
Combined	34	737	0.956	0.953	0.078

5.2.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 34 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 34 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 29 PROMIS Anxiety items served as anchor items to transform the item parameter estimates for the SF-36/MH items onto the PROMIS Anxiety 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 5.2.3 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 items were constrained to their final bank values, while the SF-36/MH items were calibrated under the constraints imposed by the anchor items.

Table 5.2.3: IRT Linking Constants

	A	B
Mean/Mean	1.068	0.166
Mean/Sigma	1.108	0.117
Haebara	1.104	0.117
Stocking-Lord	1.100	0.126

The item parameter estimates for the SF-36/MH items were linked to the PROMIS Anxiety metric using the transformation constants shown in Table 5.2.3. The SF-36/MH item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Anxiety metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for SF-36/MH as shown in Figure 5.2.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.2.6 displays the differences on the vertical axis.

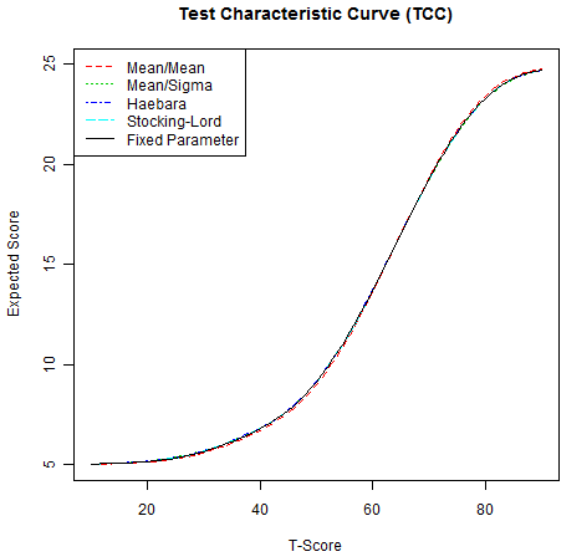

Figure 5.2.5: Test Characteristic Curves (TCC) from Different Linking Methods

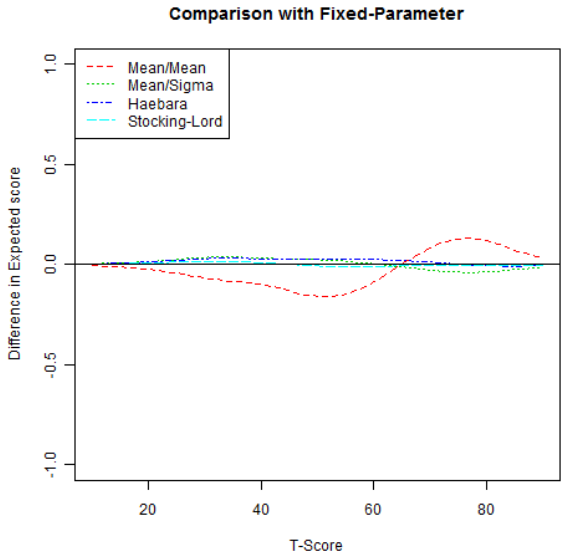

Figure 5.2.6: Difference in Test Characteristic Curves (TCC)

Table 5.2.4 shows the fixed-parameter calibration item parameter estimates for SF-36/MH. The marginal reliability estimate for SF-36/MH based on the item parameter estimates was 0.796. The marginal reliability estimates for PROMIS Anxiety and the combined set were 0.946 and 0.957, respectively. The slope parameter estimates for SF-36/MH ranged from 1.66 to 2.4 with a mean of 2.1. The slope parameter estimates for PROMIS Anxiety ranged from 1.27 to 3.88 with a mean of 2.72. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.2.7 displays the scale information functions for PROMIS Anxiety, SF-36/MH, and the combined set of 34. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.2.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.2.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
2.352	0.112	1.137	1.954	2.834
2.402	0.692	1.282	1.892	2.619
1.711	-1.619	0.315	1.107	2.237
2.374	-0.017	0.875	1.553	2.304
1.657	-1.357	0.598	1.307	2.485

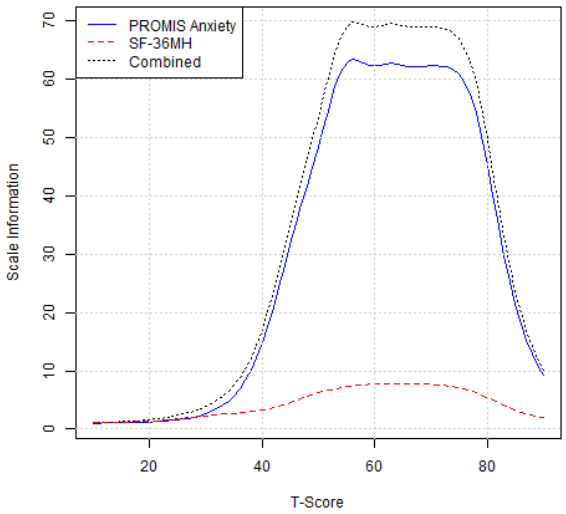

Figure 5.2.7: Comparison of Scale Information Functions

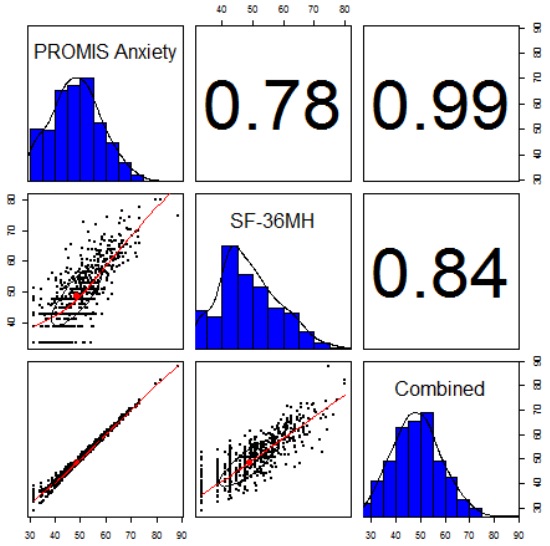

Figure 5.2.8: Comparison of IRT Scaled Scores

5.2.5. 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 SF-36/MH to a scaled score on PROMIS Anxiety can be useful. Based on the SF-36/MH item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 4 can be used to map simple raw summed scores from SF-36/MH to T-score values linked to the PROMIS Anxiety 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.

5.2.6. Equipercentile Linking

We mapped each raw summed score point on SF-36/MH to a corresponding scaled score on PROMIS Anxiety by identifying scores on PROMIS Anxiety that have the same percentile ranks as scores on SF-36/MH. 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 5.2.9 displays the cumulative distribution functions of the measures. Figure 5.2.10 shows the equipercentile linking functions based on raw summed scores, from SF-36/MH to PROMIS Anxiety. 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 5 and Appendix Table 6 show the equipercentile crosswalk tables. The result shown in Appendix Table 5 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 6 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).

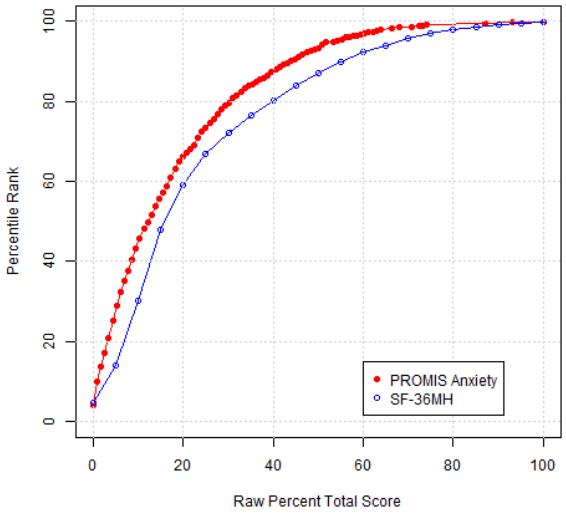

Figure 5.2.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

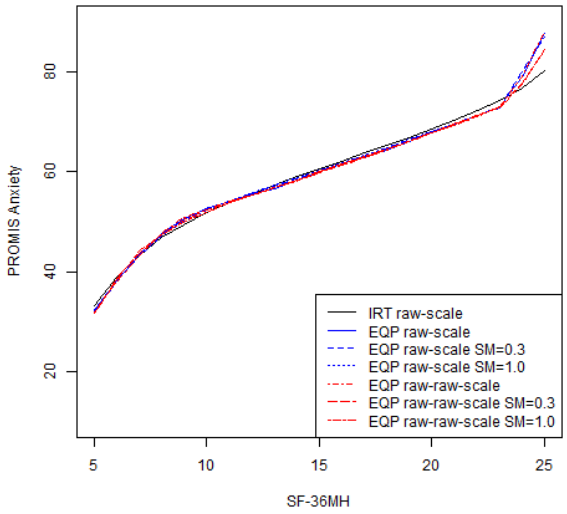

Figure 5.2.10: Equipercentile Linking Functions

5.2.7. 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 5.2.10, 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 5.2.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Anxiety T-scores and SF-36/MH scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.2.10), 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 (0.782), followed by IRT raw-scale (0.765). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (6.441), followed by IRT raw-scale (6.667).

Table 5.2.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.782	-0.241	6.441	6.441
IRT raw-scale	0.765	-0.145	6.670	6.667
EQP raw-scale SM=0.0	0.761	-0.177	6.761	6.759
EQP raw-scale SM=0.3	0.761	-0.085	6.788	6.784
EQP raw-scale SM=1.0	0.763	-0.094	6.752	6.748
EQP raw-raw-scale SM=0.0	0.761	-0.087	6.772	6.768
EQP raw-raw-scale SM=0.3	0.759	-0.102	6.769	6.765
EQP raw-raw-scale SM=1.0	0.757	-0.123	6.787	6.784

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=729), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=729) over a large number of replications (i.e., 10,000).

Table 5.2.6 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 Anxiety 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, 0.708. That is, the difference between the mean PROMIS Anxiety T-score and the mean equated SF-36/MH T-score based on a similar sample of 75 cases is expected to be around ±1.42 (i.e., 2 x 0.708).

Table 5.2.6: 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	-0.237	1.274	-0.236	0.875	-0.234	0.708
IRT raw-scale	-0.154	1.301	-0.130	0.914	-0.138	0.729
EQP raw-scale SM=0.0	-0.192	1.343	-0.187	0.918	-0.162	0.739
EQP raw-scale SM=0.3	-0.083	1.331	-0.094	0.926	-0.091	0.741
EQP raw-scale SM=1.0	-0.087	1.334	-0.100	0.916	-0.094	0.738
EQP raw-raw-scale SM=0.0	-0.073	1.341	-0.090	0.921	-0.085	0.741
EQP raw-raw-scale SM=0.3	-0.087	1.335	-0.094	0.922	-0.105	0.734
EQP raw-raw-scale SM=1.0	-0.129	1.324	-0.121	0.927	-0.123	0.751

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 4) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.3. PROMIS Depression and CES-D

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression (28 items) and CES-D. (20 items). PROMIS Depression was scaled such that higher scores represent higher levels of Depression. 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.

5.3.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 140 for PROMIS Depression and 60 for CES-D. Figure 5.3.1 and Figure 5.3.2 graphically display the raw summed score distributions of the two measures. Figure 5.3.3 shows the distribution for the combined. Figure 5.3.4 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 CES-D was 0.90. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and CES-D was 0.94. The correlations between the combined score and the measures were 0.99 and 0.95 for PROMIS Depression and CES-D, respectively.

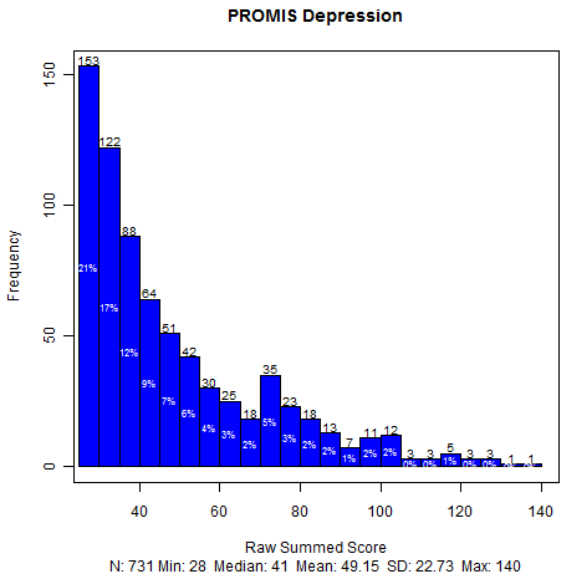

Figure 5.3.1: Raw Summed Score Distribution - PROMIS Instrument

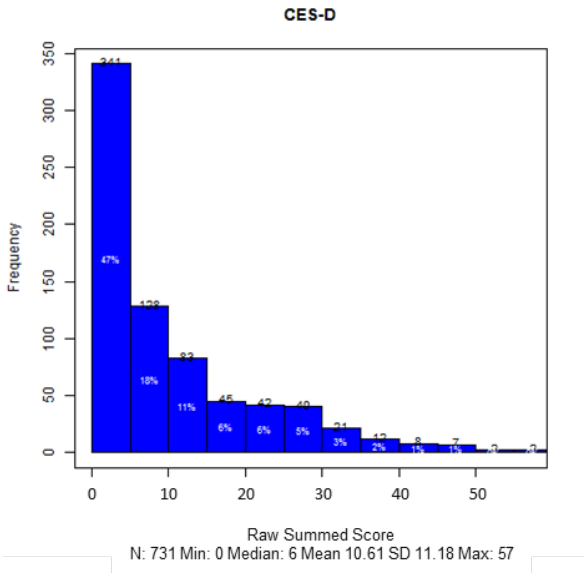

Figure 5.3.2: Raw Summed Score Distribution – Linking Instrument

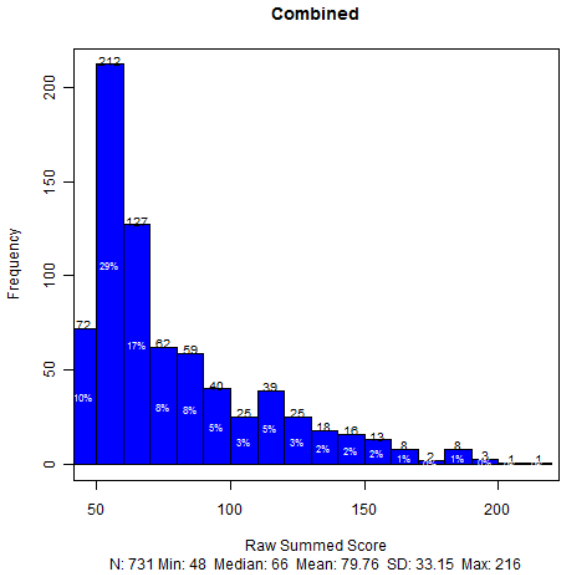

Figure 5.3.3: Raw Summed Score Distribution – Combined

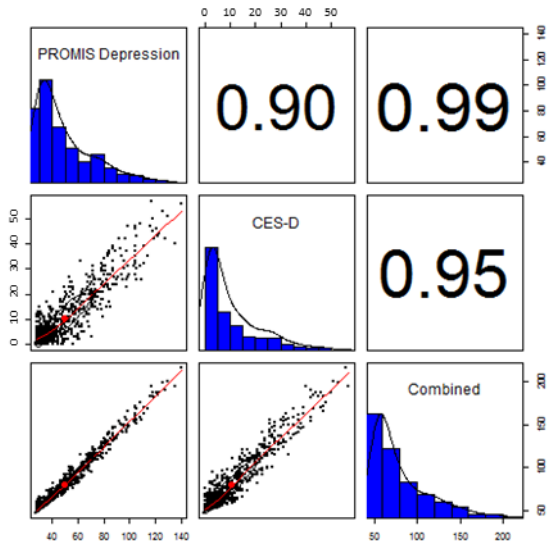

Figure 5.3.4: Scatter Plot Matrix of Raw Summed Scores

5.3.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.3.1 summarizes the results. For PROMIS Depression, Cronbach's alpha internal consistency reliability estimate was 0.980 and adjusted (corrected for overlap) item-total correlations ranged from 0.707 to 0.864. For CES-D, alpha was 0.932 and adjusted item-total correlations ranged from 0.435 to 0.823. For the 48 items, alpha was 0.982 and adjusted item-total correlations ranged from 0.446 to 0.857.

Table 5.3.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Depression	28	0.980	0.707	0.793	0.864
CES-D	20	0.932	0.435	0.628	0.823
Combined	48	0.982	0.446	0.721	0.857

5.3.3. 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 each of the two measures separately and on the combined. Table 5.3.2 summarizes the model fit statistics. For PROMIS Depression, the fit statistics were as follows: CFI = 0.984, TLI = 0.983, and RMSEA = 0.065. For CES-D, CFI = 0.942, TLI = 0.935, and RMSEA = 0.100. For the 48 items, CFI =

0.960, TLI = 0.958, and RMSEA = 0.068. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.3.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Depression	28	747	0.984	0.983	0.065
CES-D	20	747	0.942	0.935	0.100
Combined	48	747	0.960	0.958	0.068

5.3.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 48 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 48 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 28 PROMIS Depression items served as anchor items to transform the item parameter estimates for the CES-D 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 5.3.3 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 items were constrained to their final bank values, while the CES-D items were calibrated under the constraints imposed by the anchor items.

Table 5.3.3: IRT Linking Constants

	A	B
Mean/Mean	1.058	0.389
Mean/Sigma	1.119	0.339
Haebara	1.113	0.343
Stocking-Lord	1.106	0.348

The item parameter estimates for the CES-D items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.3.3. The CES-D 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 CES-D as shown in Figure 5.3.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.3.6 displays the differences on the vertical axis.

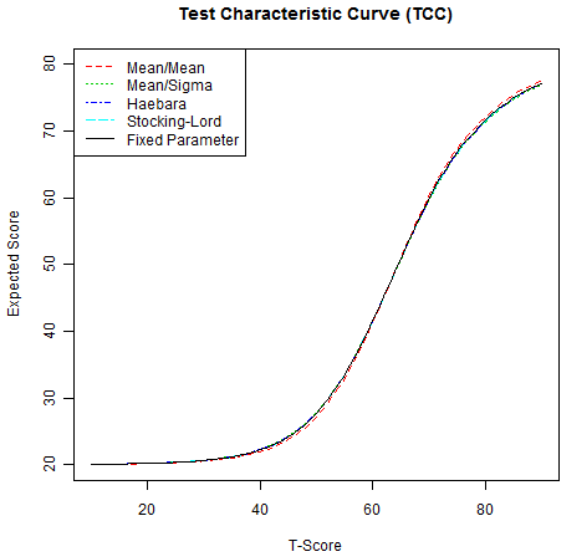

Figure 5.3.5: Test Characteristic Curves (TCC) from Different Linking Methods

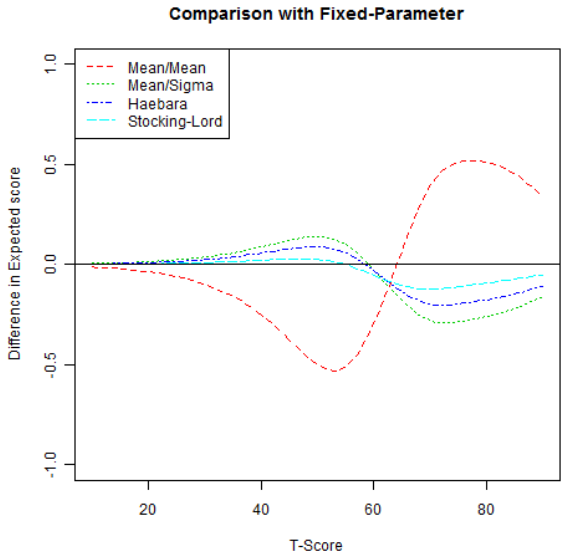

Figure 5.3.6: Difference in Test Characteristic Curves (TCC)

Table 5.3.4 shows the fixed-parameter calibration item parameter estimates for CES-D. The marginal reliability estimate for CES-D based on the item parameter estimates was 0.872. The marginal reliability estimates for PROMIS Depression and the combined set were 0.938 and 0.954, respectively. The slope parameter estimates for CES-D ranged from 1.08 to 3.63 with a mean of 1.98. The slope parameter estimates for PROMIS Depression ranged from 2.02 to 4.45 with a mean of 3.14. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.3.7 displays the scale information functions for PROMIS Depression, CES-D, and the combined set of 48. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.3.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.3.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3
2.072	0.877	1.923	3.067
1.261	1.389	2.672	3.725
3.512	0.834	1.317	1.951
1.117	0.650	1.381	2.084
1.603	0.430	1.528	2.727
3.634	0.494	1.177	1.731
1.826	0.288	1.369	2.136
1.340	-0.066	0.824	1.622
3.002	0.748	1.375	1.857
2.059	1.173	2.044	3.271
1.075	-0.463	0.949	2.162
2.228	0.169	0.946	1.738
1.287	0.343	1.698	2.919
2.175	0.492	1.293	1.866
1.395	0.966	2.323	3.608
2.131	0.273	0.923	1.810
1.717	1.609	2.320	3.474
2.810	0.262	1.250	1.986
1.833	0.785	1.877	2.641
1.490	-0.140	1.257	2.299

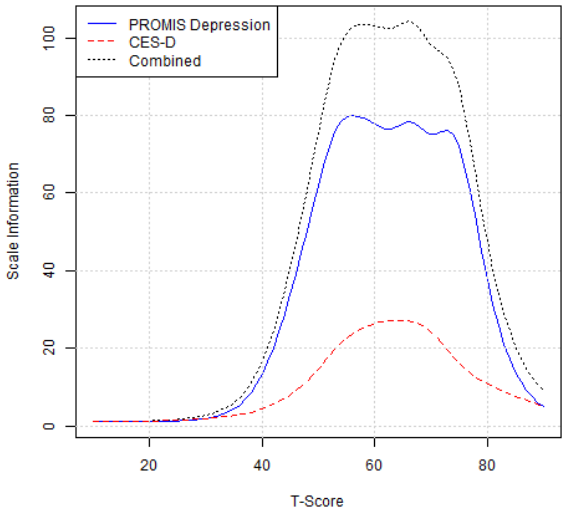

Figure 5.3.7: Comparison of Scale Information Functions

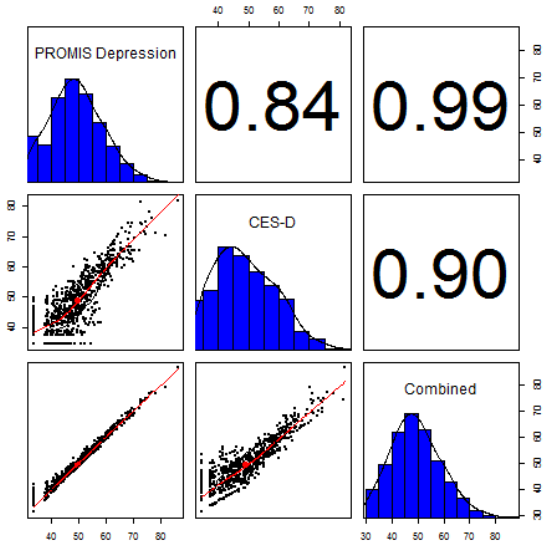

Figure 5.3.8: Comparison of IRT Scaled Scores

5.3.5. 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 CES-D to a scaled score on PROMIS Depression can be useful. Based on the CES-D item parameters derived from the fixed-

parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 7 can be used to map simple raw summed scores from CES-D 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.

5.3.6. Equipercentile Linking

We mapped each raw summed score point on CES-D to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on CES-D. 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 5.3.9 displays the cumulative distribution functions of the measures. Figure 5.3.10 shows the equipercentile linking functions based on raw summed scores, from CES-D 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 8 and Appendix Table 9 show the equipercentile crosswalk tables. The result shown in Appendix Table 8 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 9 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).

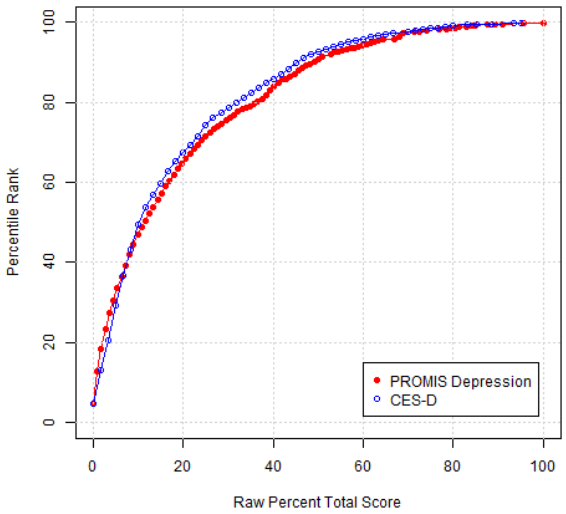

Figure 5.3.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

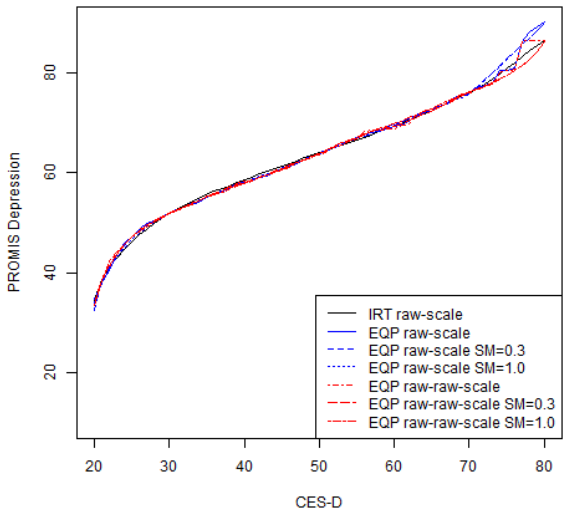

Figure 5.3.10: Equipercentile Linking Functions

5.3.7. 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 5.3.10, 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 5.3.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Depression T-scores and CES-D scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.3.10), 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 (0.844), followed by IRT raw-scale (0.821). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.461), followed by IRT raw-scale (5.772).

Table 5.3.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.844	0.307	5.456	5.461
IRT raw-scale	0.821	0.090	5.775	5.772
EQP raw-scale SM=0.0	0.813	0.054	5.896	5.892
EQP raw-scale SM=0.3	0.810	0.216	6.032	6.031
EQP raw-scale SM=1.0	0.819	0.124	5.822	5.820
EQP raw-raw-scale SM=0.0	0.812	0.034	5.928	5.924
EQP raw-raw-scale SM=0.3	0.814	0.035	5.886	5.882
EQP raw-raw-scale SM=1.0	0.815	-0.047	5.836	5.832

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=731), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=731) over a large number of replications (i.e., 10,000).

Table 5.3.6 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, 0.599. That is, the difference between the mean PROMIS Depression T-score and the mean equated CES-D T-score based on a similar sample of 75 cases is expected to be around ±1.20 (i.e., 2 x 0.599).

Table 5.3.6: 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	0.293	1.060	0.301	0.744	0.303	0.599
IRT raw-scale	0.074	1.136	0.100	0.781	0.084	0.633
EQP raw-scale SM=0.0	0.049	1.166	0.056	0.806	0.063	0.652
EQP raw-scale SM=0.3	0.217	1.177	0.225	0.825	0.219	0.658
EQP raw-scale SM=1.0	0.126	1.145	0.124	0.788	0.121	0.634
EQP raw-raw-scale SM=0.0	0.054	1.174	0.043	0.806	0.036	0.649
EQP raw-raw-scale SM=0.3	0.042	1.172	0.036	0.801	0.028	0.645
EQP raw-raw-scale SM=1.0	-0.058	1.153	-0.047	0.793	-0.045	0.639

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 7) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.4. PROMIS Depression and SF-36/MH

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression (28 items) and SF-36/MH (5 items). PROMIS Depression was scaled such that higher scores represent higher levels of Depression; for the SF-36/MH, higher scores represent lower levels of depression. 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.

5.4.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 140 for PROMIS Depression and 25 for SF-36/MH. Figure 5.4.1 and Figure 5.4.2 graphically display the raw summed score distributions of the two measures. Figure 5.4.3 shows the distribution for the combined. Figure 5.4.4 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 SF-36/MH was -0.86. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and SF-36/MH was -0.93. The correlations between the combined score and the measures were 1.00 and 0.90 for PROMIS Depression and SF-36/MH, respectively.

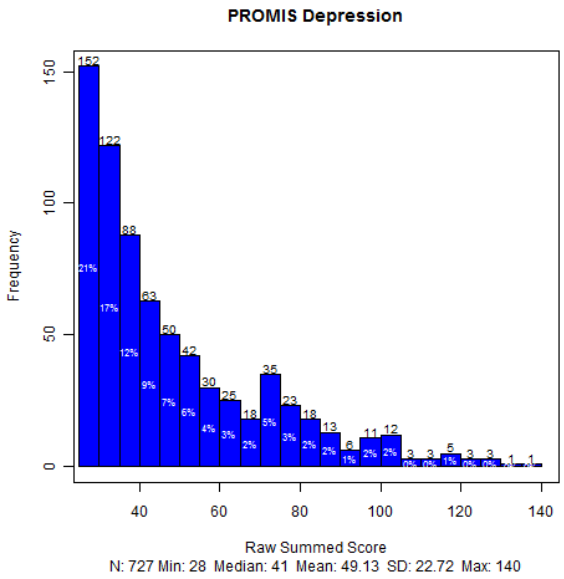

Figure 5.4.1: Raw Summed Score Distribution - PROMIS Instrument

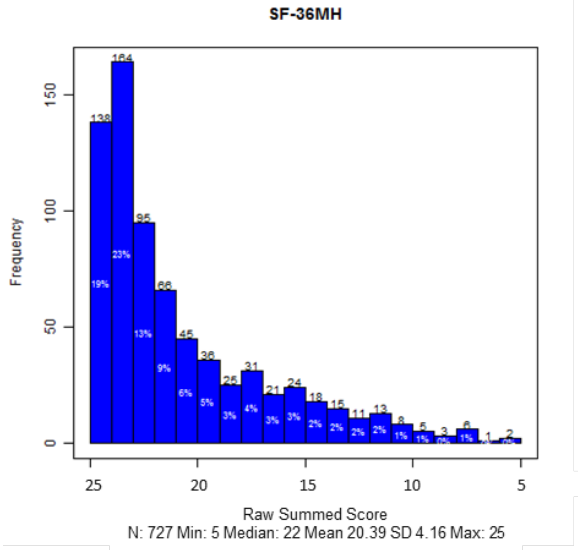

Figure 5.4.2: Raw Summed Score Distribution – Linking Instrument

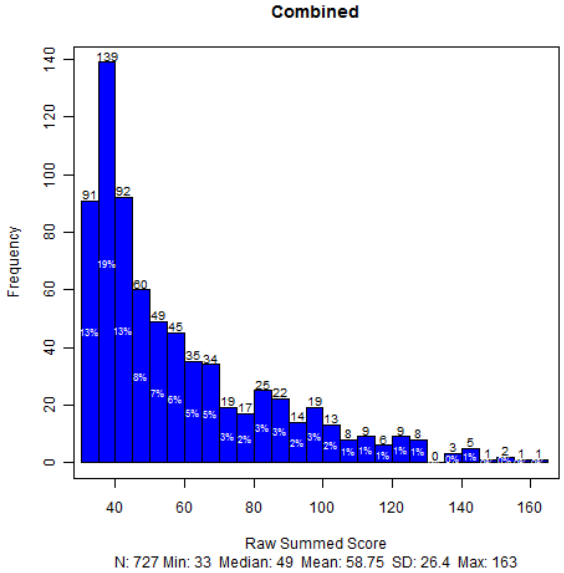

Figure 5.4.3: Raw Summed Score Distribution – Combined

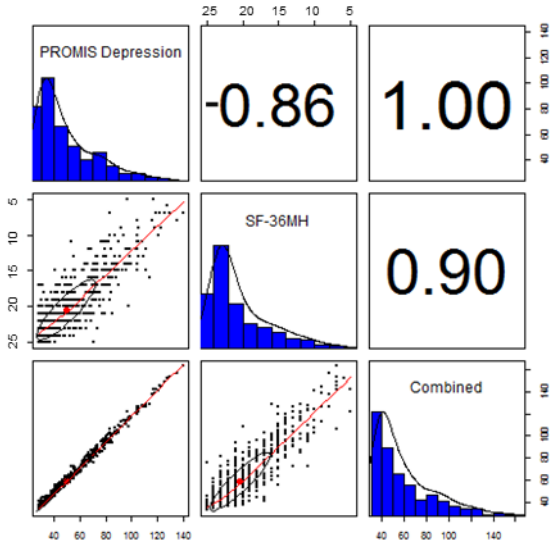

Figure 5.4.4: Scatter Plot Matrix of Raw Summed Scores

5.4.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.4.1 summarizes the results. For PROMIS Depression, Cronbach's alpha internal consistency reliability estimate was 0.980 and adjusted (corrected for overlap) item-total correlations ranged from 0.707 to 0.863. For SF-36/MH, alpha was 0.888 and adjusted item-total correlations ranged from 0.578 to 0.818. For the 33 items, alpha was 0.982 and adjusted item-total correlations ranged from 0.621 to 0.862.

Table 5.4.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Depression	28	0.980	0.707	0.792	0.863
SF-36/MH	5	0.888	0.578	0.731	0.818
Combined	33	0.982	0.621	0.782	0.862

5.4.3. 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 each of the two measures separately and on the combined. Table 5.4.2 summarizes the model fit statistics. For PROMIS Depression, the fit statistics were as follows: CFI = 0.984, TLI = 0.983, and RMSEA = 0.064. For SF-36/MH, CFI = 0.988, TLI = 0.976, and RMSEA = 0.207. For the 33 items, CFI =

0.975, TLI = 0.973, and RMSEA = 0.074. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.4.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Depression	28	737	0.984	0.983	0.064
SF-36/MH	5	737	0.988	0.976	0.207
Combined	33	737	0.975	0.973	0.074

5.4.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 33 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 33 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 28 PROMIS Depression items served as anchor items to transform the item parameter estimates for the SF-36/MH 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 5.4.3 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 items were constrained to their final bank values, while the SF-36/MH items were calibrated under the constraints imposed by the anchor items.

Table 5.4.3: IRT Linking Constants

	A	B
Mean/Mean	1.023	0.266
Mean/Sigma	1.075	0.215
Haebara	1.070	0.220
Stocking-Lord	1.064	0.225

The item parameter estimates for the SF-36/MH items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.4.3. The SF-36/MH 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 SF-36/MH as shown in Figure 5.4.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.4.6 displays the differences on the vertical axis.

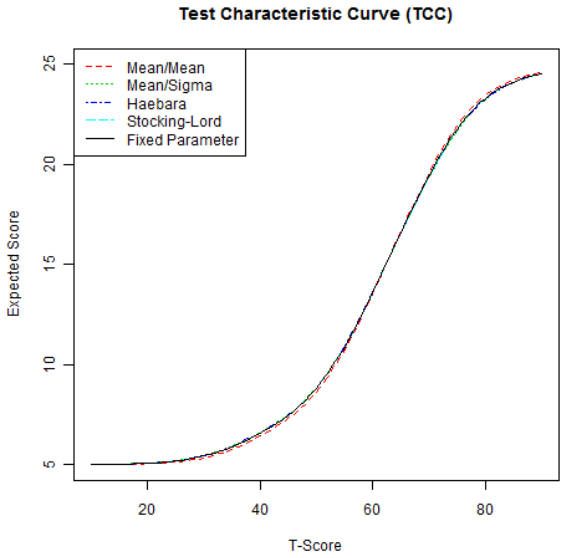

Figure 5.4.5: Test Characteristic Curves (TCC) from Different Linking Methods

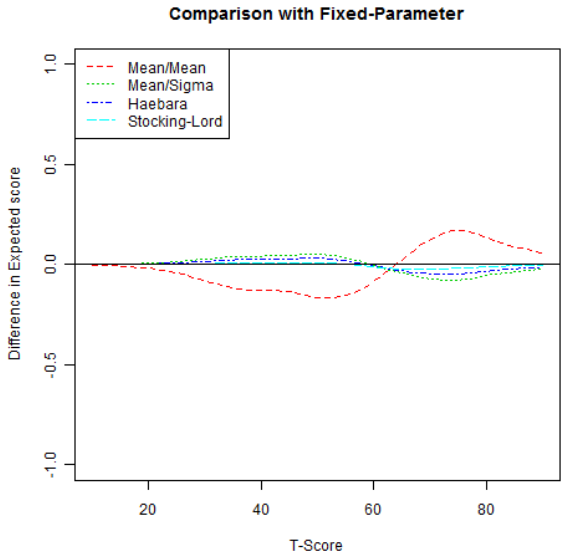

Figure 5.4.6: Difference in Test Characteristic Curves (TCC)

Table 5.4.4 shows the fixed-parameter calibration item parameter estimates for SF-36/MH. The marginal reliability estimate for SF-36/MH based on the item parameter estimates was 0.826. The marginal reliability estimates for PROMIS Depression and the combined set were 0.938 and 0.955, respectively. The slope parameter estimates for SF-36/MH ranged from 1.62 to 3.35 with a mean of 2.46. The slope parameter estimates for PROMIS Depression ranged from 2.02 to 4.45 with a mean of 3.14. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.4.7 displays the scale information functions for PROMIS Depression, SF-36/MH, and the combined set of 33. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.4.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.4.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
1.615	0.225	1.430	2.404	3.443
3.229	0.664	1.205	1.806	2.507
1.921	-1.423	0.357	1.119	2.196
3.354	0.028	0.837	1.473	2.176
2.189	-1.112	0.568	1.219	2.264

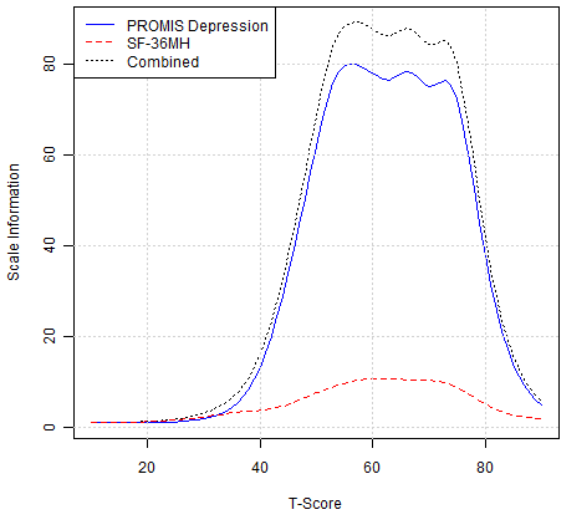

Figure 5.4.7: Comparison of Scale Information Functions

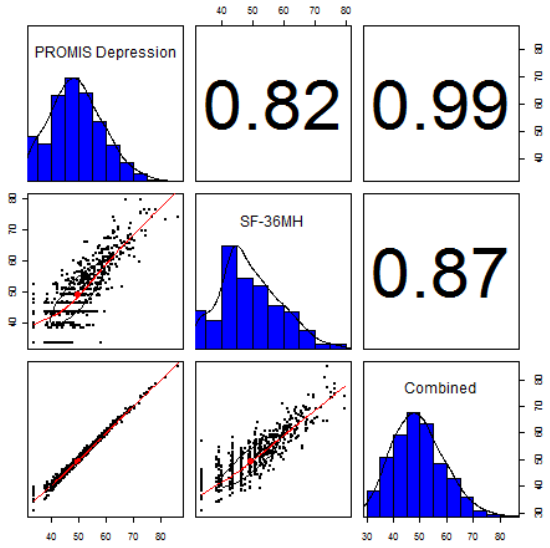

Figure 5.4.8: Comparison of IRT Scaled Scores

5.4.5. 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 SF-36/MH to a scaled score on PROMIS Depression can be useful. Based on the SF-36/MH item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 10 can be used to map simple raw summed scores from SF-36/MH 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.

5.4.6. Equipercentile Linking

We mapped each raw summed score point on SF-36/MH to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on SF-36/MH. 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 5.4.9 displays the cumulative distribution functions of the measures. Figure 5.4.10 shows the equipercentile linking functions based on raw summed scores, from SF-36/MH 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 11 and Appendix Table 12 show the equipercentile crosswalk tables. The result shown in Appendix Table 11 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 12 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).

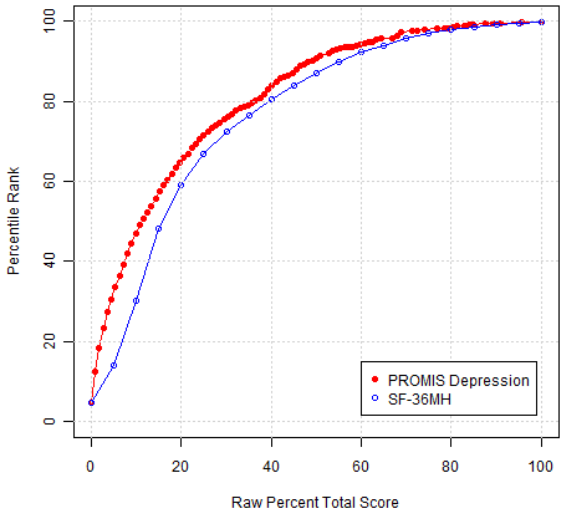

Figure 5.4.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

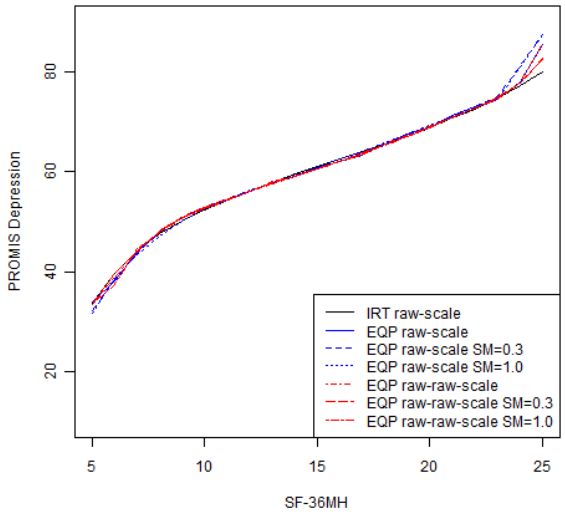

Figure 5.4.10: Equipercentile Linking Functions

5.4.7. 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 5.4.10, 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 5.4.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Depression T-

scores and SF-36/MH scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.4.10), 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 (0.819), followed by IRT raw-scale (0.806). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.791), followed by IRT raw-scale (5.961).

Table 5.4.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.819	0.034	5.795	5.791
IRT raw-scale	0.806	0.070	5.965	5.961
EQP raw-scale SM=0.0	0.803	0.037	6.042	6.038
EQP raw-scale SM=0.3	0.800	0.334	6.217	6.222
EQP raw-scale SM=1.0	0.801	0.489	6.250	6.265
EQP raw-raw-scale SM=0.0	0.802	0.036	6.060	6.056
EQP raw-raw-scale SM=0.3	0.804	-0.034	6.000	5.996
EQP raw-raw-scale SM=1.0	0.800	0.108	6.106	6.103

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=727), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=727) over a large number of replications (i.e., 10,000).

Table 5.4.6 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, 0.632. That is, the difference between the mean PROMIS Depression T-score and the mean equated SF-36/MH T-score based on a similar sample of 75 cases is expected to be around ±1.26 (i.e., 2 x 0.632).

Table 5.4.6: 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	0.014	1.127	0.035	0.796	0.034	0.632
IRT raw-scale	0.063	1.164	0.065	0.826	0.070	0.658
EQP raw-scale SM=0.0	0.065	1.187	0.012	0.824	0.021	0.661
EQP raw-scale SM=0.3	0.348	1.216	0.358	0.857	0.341	0.682
EQP raw-scale SM=1.0	0.460	1.231	0.486	0.846	0.494	0.682
EQP raw-raw-scale SM=0.0	0.021	1.187	0.036	0.817	0.029	0.660
EQP raw-raw-scale SM=0.3	-0.034	1.194	-0.036	0.822	-0.032	0.650
EQP raw-raw-scale SM=1.0	0.107	1.207	0.113	0.829	0.117	0.653

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 10) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.5. PROMIS Anger and BPAQ

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Anger, namely the PROMIS Anger (29 items) and BPAQ (12 items). PROMIS Anger was scaled such that higher scores represent higher levels of Anger. 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.

5.5.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 145 for PROMIS Anger and 60 for BPAQ. Figure 5.5.1 and Figure 5.5.2 graphically display the raw summed score distributions of the two measures. Figure 5.5.3 shows the distribution for the combined. Figure 5.5.4 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 Anger and Aggression Questionnaire was 0.59. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anger and BPAQ was 0.65. The correlations between the combined score and the measures were 0.95 and 0.80 for PROMIS Anger and BPAQ, respectively.

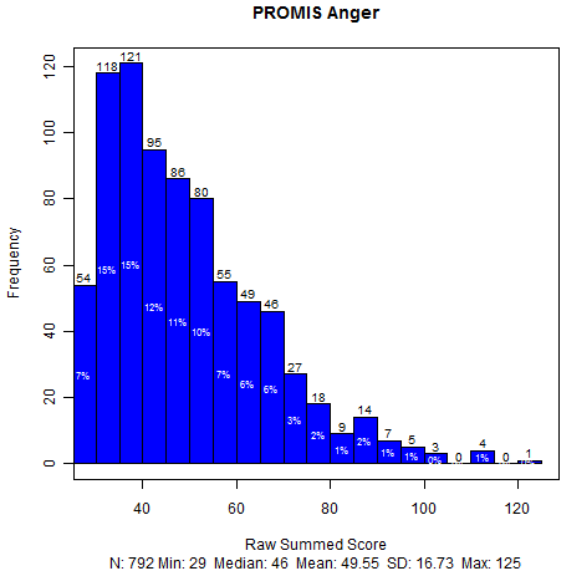

Figure 5.5.1: Raw Summed Score Distribution - PROMIS Instrument

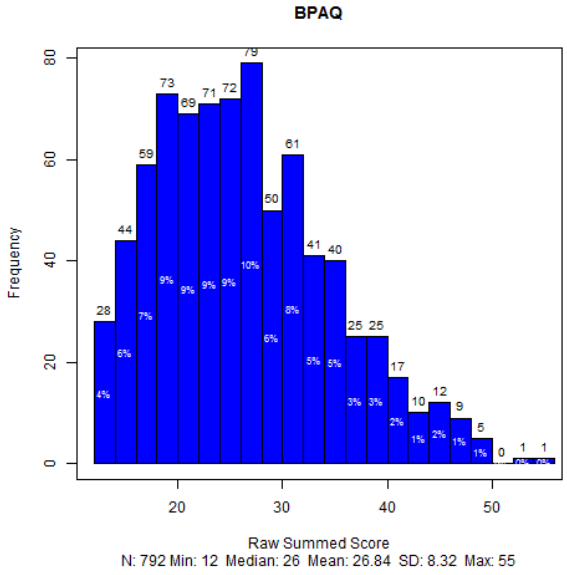

Figure 5.5.2: Raw Summed Score Distribution – Linking Instrument

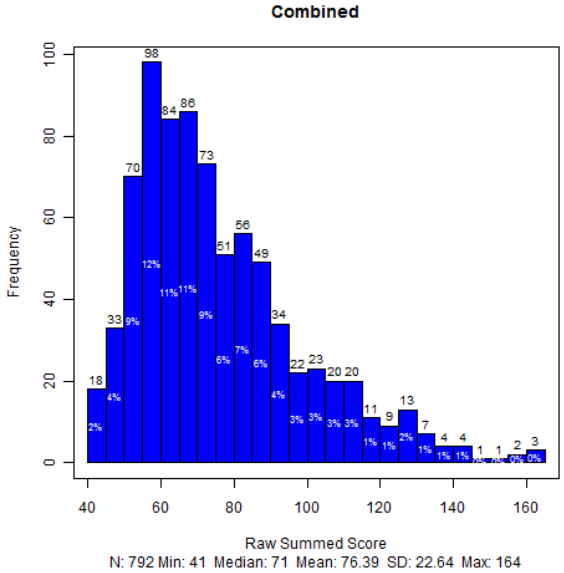

Figure 5.5.3: Raw Summed Score Distribution – Combined

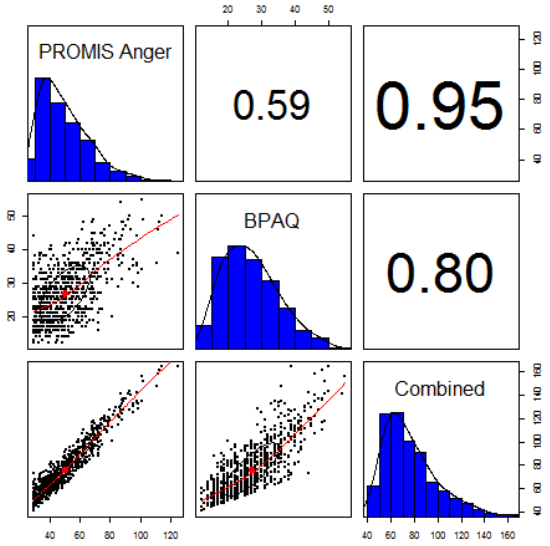

Figure 5.5.4: Scatter Plot Matrix of Raw Summed Scores

5.5.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.5.1 summarizes the results. For PROMIS Anger, Cronbach's alpha internal consistency reliability estimate was 0.957 and adjusted (corrected for overlap) item-total correlations ranged from 0.489 to 0.759. For Aggression Questionnaire, alpha was 0.848 and adjusted item-total correlations ranged from 0.270 to 0.645. For the 41 items, alpha was 0.952 and adjusted item-total correlations ranged from 0.119 to 0.749.

Table 5.5.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Anger	29	0.957	0.489	0.648	0.759
BPAQ	12	0.848	0.270	0.522	0.645
Combined	41	0.952	0.119	0.577	0.749

5.5.3. 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 each of the two measures separately and on the combined. Table 5.5.2 summarizes the model fit statistics. For PROMIS Anger, the fit statistics were as follows: CFI = 0.973, TLI = 0.971, and RMSEA= 0.052. For BPAQ, CFI = 0.905, TLI = 0.884, and RMSEA= 0.13. For the 41 items, CFI = 0.892, TLI =

0.886, and RMSEA = 0.081. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.5.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Anger	29	824	0.973	0.971	0.052
BPAQ	12	824	0.905	0.884	0.130
Combined	41	824	0.892	0.886	0.081

5.5.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 41 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 41 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 29 PROMIS Anger items served as anchor items to transform the item parameter estimates for the BPAQ items onto the PROMIS Anger 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 5.5.3 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 items were constrained to their final bank values, while the BPAQ items were calibrated under the constraints imposed by the anchor items.

Table 5.5.3: IRT Linking Constants

	A	B
Mean/Mean	0.993	-0.180
Mean/Sigma	0.996	-0.186
Haebara	0.983	-0.167
Stocking-Lord	0.992	-0.179

The item parameter estimates for the BPAQ items were linked to the PROMIS Anger metric using the transformation constants shown in Table 5.5.3. The BPAQ item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Anger metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for BPAQ as shown in Figure 5.5.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.5.6 displays the differences on the vertical axis.

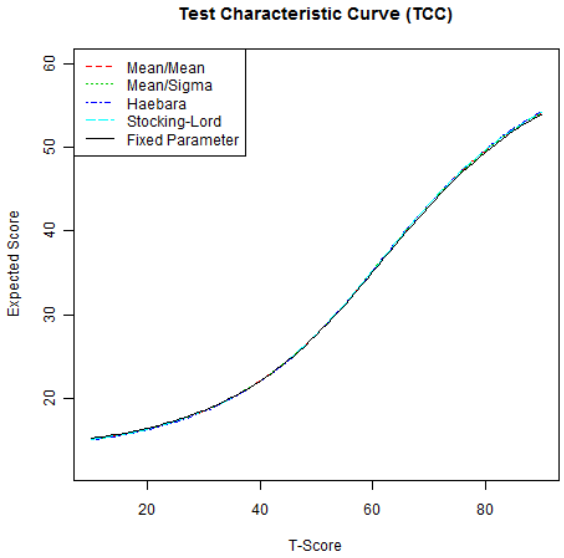

Figure 5.5.5: Test Characteristic Curves (TCC) from Different Linking Methods

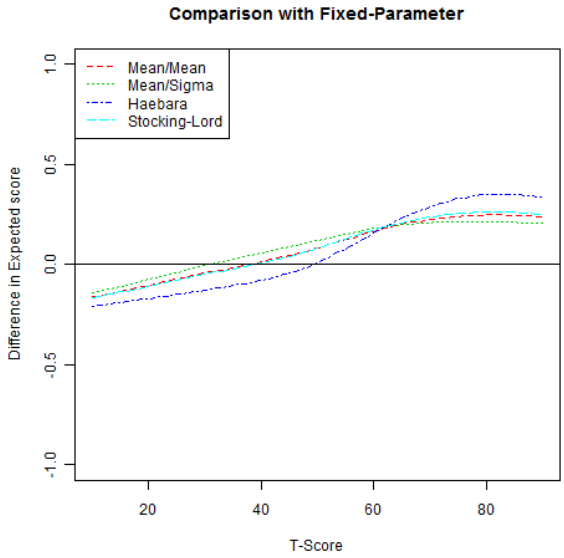

Figure 5.5.6: Difference in Test Characteristic Curves (TCC)

Table 5.5.4 shows the fixed-parameter calibration item parameter estimates for BPAQ. The marginal reliability estimate for BPAQ based on the item parameter estimates was 0.782. The marginal reliability estimates for PROMIS Anger and the combined set were 0.955 and 0.962, respectively. The slope parameter estimates for BPAQ ranged from 0.215 to 1.76 with a mean of 1.07. The slope parameter estimates for PROMIS Anger ranged from 1.41 to 2.99 with a mean of 2.21. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.5.7 displays the scale information functions for PROMIS Anger, BPAQ, and the combined set of 41. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.5.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.5.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
1.271	0.376	1.017	1.966	3.781
0.215	-13.585	-6.481	-2.259	6.485
0.925	-0.633	0.636	2.020	4.059
1.137	-0.760	1.005	1.914	3.049
0.664	-1.996	-0.651	1.057	3.822
0.550	-1.916	0.025	1.574	4.971
1.590	0.079	0.853	1.578	2.786
1.139	-2.004	-0.669	0.055	2.718
0.935	-1.730	-0.223	1.496	3.967
1.757	0.541	0.990	1.642	2.776
1.084	-0.423	0.402	1.372	3.643
1.563	0.557	1.154	1.996	3.091

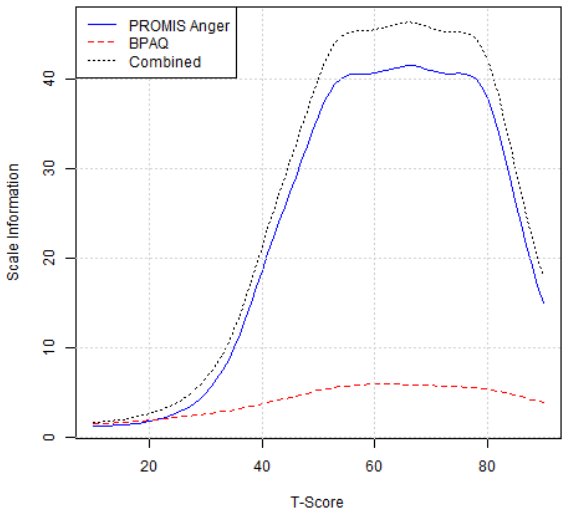

Figure 5.5.7: Comparison of Scale Information Functions

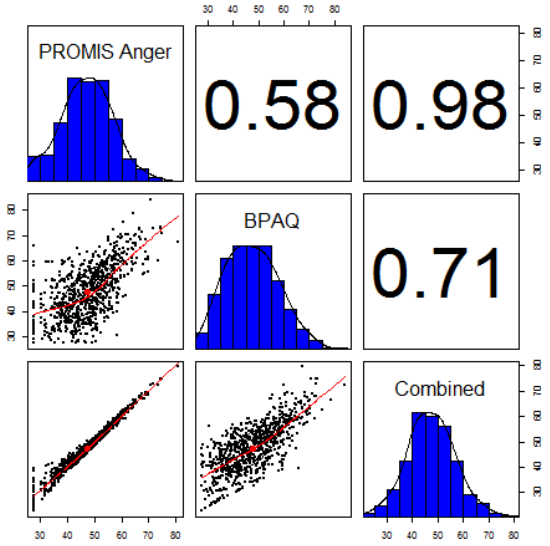

Figure 5.5.8: Comparison of IRT Scaled Scores

5.5.5. 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 BPAQ to a scaled score on PROMIS Anger can be useful. Based on the BPAQ item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 13 can be used to map simple raw summed scores from BPAQ to T-score values linked to the PROMIS Anger 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.

5.5.6. Equipercentile Linking

We mapped each raw summed score point on BPAQ to a corresponding scaled score on PROMIS Anger by identifying scores on PROMIS Anger that have the same percentile ranks as scores on BPAQ. 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 5.5.9 displays the cumulative distribution functions of the measures. Figure 5.5.10 shows the equipercentile linking functions based on raw summed scores, from BPAQ to PROMIS Anger.

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 14 and Appendix Table 15 show the equipercentile crosswalk tables. The result shown in Appendix Table 14 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 15 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).

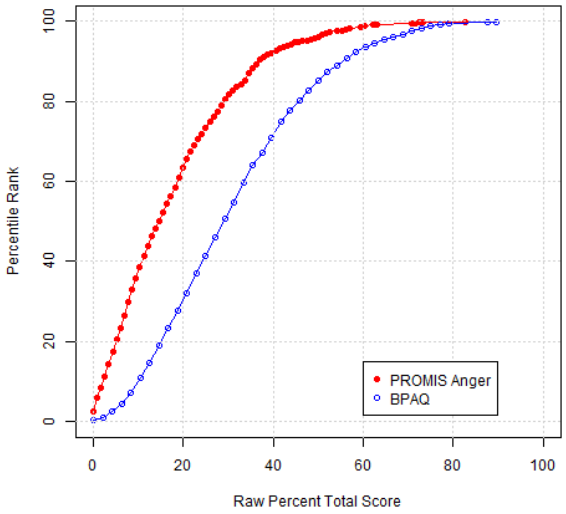

Figure 5.5.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

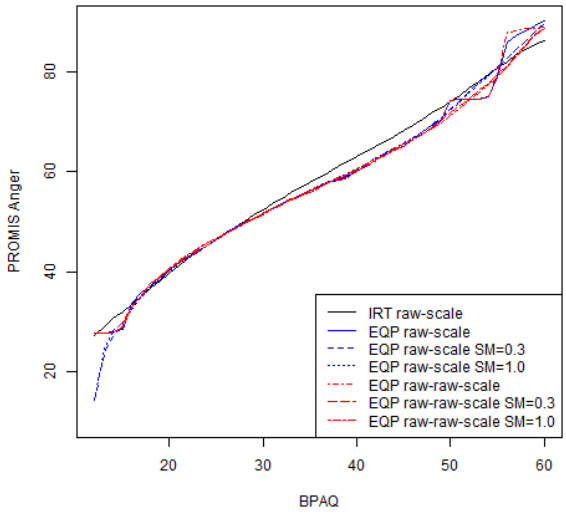

Figure 5.5.10: Equipercentile Linking Functions

5.5.7. 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 5.5.10, 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 5.5.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Anger T-scores and BPAQ scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.5.10), 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 (0.585), followed by IRT raw-scale (0.534). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (9.061), followed by EQP raw-raw-scale SM=0.3 (9.16). The low correlations indicate the two measures may be significantly different from each other. The disattenuated correlation of 0.65 was still very low (less than 0.80). Caution should be demonstrated when using these linking tables.

Table 5.5.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.585	-0.137	9.066	9.061
IRT raw-scale	0.534	-0.602	9.450	9.463
EQP raw-scale SM=0.0	0.526	0.036	9.191	9.186
EQP raw-scale SM=0.3	0.521	0.179	9.462	9.458
EQP raw-scale SM=1.0	0.524	0.194	9.461	9.457
EQP raw-raw-scale SM=0.0	0.526	0.035	9.188	9.182
EQP raw-raw-scale SM=0.3	0.526	-0.040	9.165	9.160
EQP raw-raw-scale SM=1.0	0.524	-0.067	9.178	9.172

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=792), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=792) over a large number of replications (i.e., 10,000).

Table 5.5.6 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 Anger 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, 0.994. That is, the difference between the mean PROMIS Anger T-score and the mean equated BPAQ T-score based on a similar sample of 75 cases is expected to be around ±1.99 (i.e., 2 x 0.994).

Table 5.5.6: 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	-0.144	1.807	-0.133	1.251	-0.136	0.994
IRT raw-scale	-0.627	1.879	-0.613	1.300	-0.638	1.030
EQP raw-scale SM=0.0	0.011	1.831	0.032	1.257	0.031	0.997
EQP raw-scale SM=0.3	0.167	1.860	0.178	1.291	0.175	1.040
EQP raw-scale SM=1.0	0.187	1.855	0.178	1.291	0.191	1.048
EQP raw-raw-scale SM=0.0	0.047	1.803	0.034	1.266	0.048	1.005
EQP raw-raw-scale SM=0.3	-0.039	1.807	-0.043	1.262	-0.041	1.009
EQP raw-raw-scale SM=1.0	-0.055	1.807	-0.062	1.236	-0.056	1.016

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 13) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.6. PROMIS Physical Function and HAQ-DI

Note: This linking analysis has been revised since its initial publication in Volume 1 (2012). Interested readers may obtain updated results on the Linking Tables and Publications sections of the prosettastone.org website:

http://www.prosettastone.org/LinkingTables1/Pages/default.aspx

http://www.prosettastone.org/PublicationsPresentations/Pages/default.aspx

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Physical Function, namely the PROMIS Physical Function (76 items) and HAQ-DI (20 items). Both PROMIS Physical Function and the HAQ-DI were scaled such that higher scores represent higher levels of Physical Function. 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.

5.6.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 380 for PROMIS Physical Function and 73 for HAQ-DI. Figure 5.6.1 and Figure 5.6.2 graphically display the raw summed score distributions of the two measures. Figure 5.6.3 shows the distribution for the combined. Figure 5.6.4 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 Physical Function and HAQ-DI was 0.91. The disattenuated (corrected for unreliabilities) correlation between PROMIS Physical Function and HAQ was 0.95. The correlations between the combined score and the measures were 1.00 and 0.93 for PROMIS Physical Function and HAQ-DI, respectively.

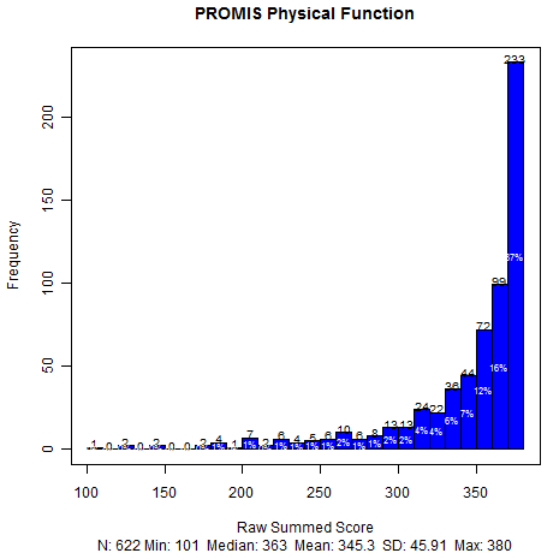

Figure 5.6.1: Raw Summed Score Distribution - PROMIS Instrument

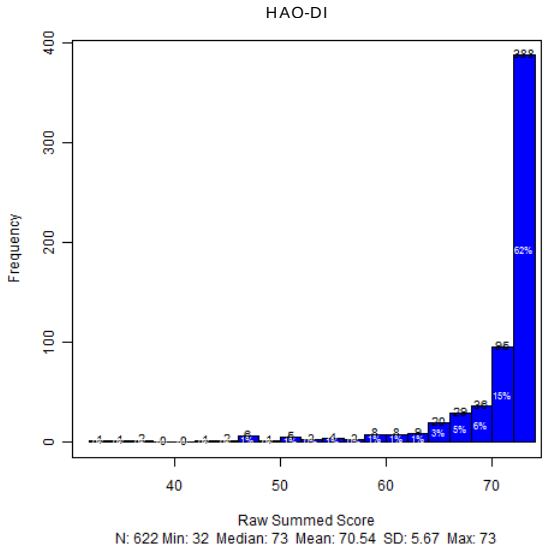

Figure 5.6.2: Raw Summed Score Distribution – Linking Instrument

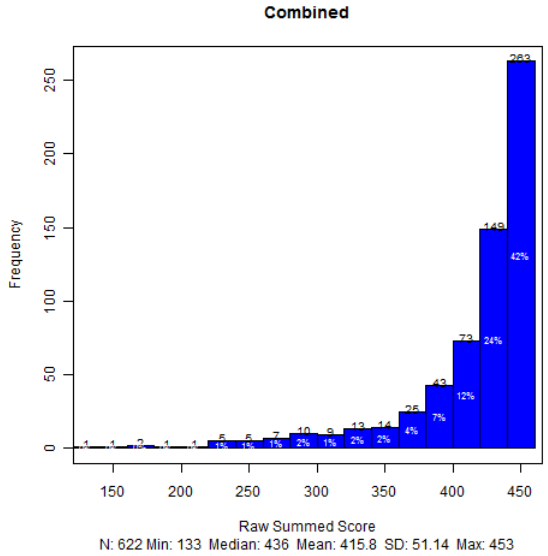

Figure 5.6.3: Raw Summed Score Distribution – Combined

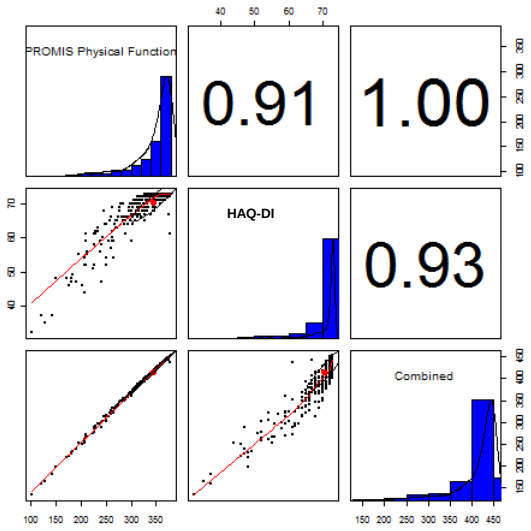

Figure 5.6.4: Scatter Plot Matrix of Raw Summed Scores

5.6.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.6.1 summarizes the results. For PROMIS Physical Function, Cronbach's alpha internal consistency reliability estimate was 0.987 and adjusted (corrected for overlap) item-total correlations ranged from 0.517 to 0.868. For HAQ-DI, alpha was 0.94 and adjusted item-total correlations ranged from 0.441 to 0.772. For the 96 items, alpha was 0.988 and adjusted item-total correlations ranged from 0.374 to 0.865.

Table 5.6.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Physical Function	76	0.987	0.517	0.729	0.868
HAQ-DI	20	0.940	0.441	0.671	0.772
Combined	96	0.988	0.374	0.713	0.865

5.6.3. 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 each of the two measures separately and on the combined. Table 5.6.2 summarizes the model fit statistics. For PROMIS Physical Function, the fit statistics were as follows: CFI = 0.977, TLI = 0.976, and RMSEA = 0.043. For HAQ-DI, CFI = 0.983, TLI = 0.981, and RMSEA = 0.039. For the 96 items, CFI = 0.973, TLI = 0.972, and RMSEA = 0.037. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.6.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Physical Function	76	722	0.977	0.976	0.043
HAQ-DI	20	722	0.983	0.981	0.039
Combined	96	722	0.973	0.972	0.037

5.6.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 96 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 96 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 76 PROMIS Physical Function items served as anchor items to transform the item parameter estimates for the HAQ-DI items onto the PROMIS Physical Function 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 5.6.3 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 items were constrained to their final bank values, while the HAQ-DI items were calibrated under the constraints imposed by the anchor items.

Table 5.6.3: IRT Linking Constants

	A	B
Mean/Mean	1.540	-1.643
Mean/Sigma	1.591	-1.642
Haebara	1.564	-1.633
Stocking-Lord	1.581	-1.630

The item parameter estimates for the HAQ-DI items were linked to the PROMIS Physical Function metric using the transformation constants shown in Table 5.6.3. The HAQ-DI item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Physical Function metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for HAQ-DI as shown in Figure 5.6.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.6.6 displays the differences on the vertical axis.

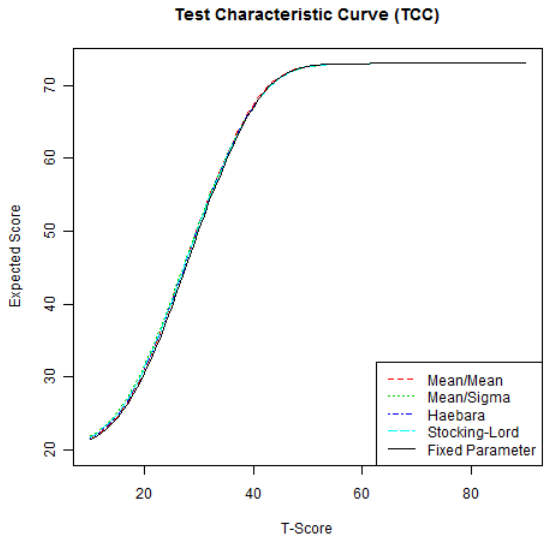

Figure 5.6.5: Test Characteristic Curves (TCC) from Different Linking Methods

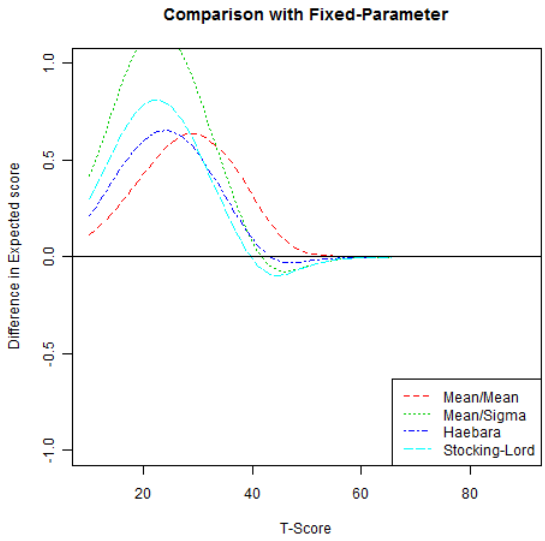

Figure 5.6.6: Difference in Test Characteristic Curves (TCC)

Table 5.6.4 shows the fixed-parameter calibration item parameter estimates for HAQ-DI. The marginal reliability estimate for HAQ-DI based on the item parameter estimates 0.644. The marginal reliability estimates for PROMIS Physical Function and the combined set were 0.956 and 0.957, respectively. The slope parameter estimates for HAQ-DI ranged from 2.23 to 4.41 with a mean of 3.82. The slope parameter estimates for PROMIS Physical Function ranged from 2.03 to 4.83 with a mean of 3.29. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.6.7 displays the scale information functions for PROMIS Physical Function, HAQ-DI, and the combined set of 96. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.6.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.6.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3
3.612	-2.829	-2.251	-1.322
2.472	-3.403	-3.102	-1.904
2.875	-2.826	-2.099	-0.989
2.564	-2.745	-1.543	
3.013	-3.037	-2.660	-2.076
2.368	-2.602		
2.331	-3.453	-3.265	-1.723
3.769	-2.537	-2.203	-1.388
3.641	-2.634	-2.013	-1.090
3.609	-3.196	-2.660	-1.679
3.210	-1.844	-1.594	-0.975
2.502	-2.597	-1.691	
3.246	-2.369	-1.950	-1.013
3.019	-3.035	-2.091	-1.085
3.125	-2.902	-1.963	
2.230	-3.270	-1.816	

2.635	-3.151	-2.251	
3.944	-2.464	-1.908	-1.138
2.993	-3.427	-2.471	-1.261
4.413	-1.942	-1.441	-0.574

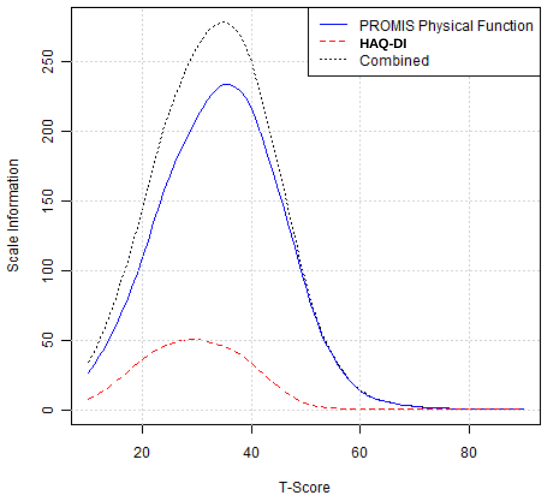

Figure 5.6.7: Comparison of Scale Information Functions

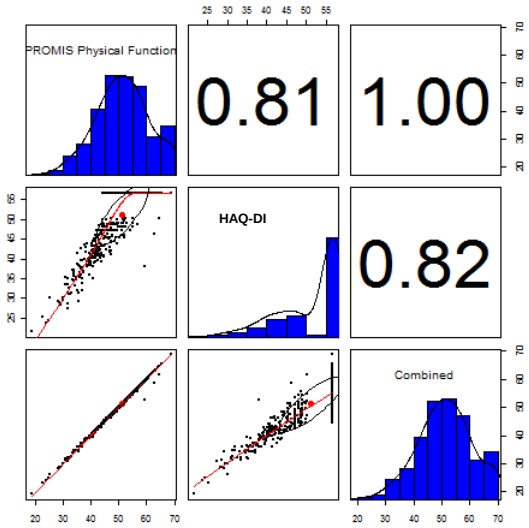

Figure 5.6.8: Comparison of IRT Scaled Scores

5.6.5. 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 HAQ-DI to a scaled score on PROMIS Physical Function can be useful. Based on the HAQ-DI item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 16 can be used to map simple raw summed scores from HAQ-DI to T-score values linked to the PROMIS Physical Function 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.

5.6.6. Equipercentile Linking

We mapped each raw summed score point on HAQ-DI to a corresponding scaled score on PROMIS Physical Function by identifying scores on PROMIS Physical Function that have the same percentile ranks as scores on HAQ-DI. 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 5.1.9 displays the cumulative distribution functions of the measures. Figure 5.1.10 shows the equipercentile linking functions based on raw summed scores, from HAQ-DI to PROMIS Physical Function. 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 17 and Appendix Table 18 show the equipercentile crosswalk tables. The result shown in Appendix Table 17 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 18 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).

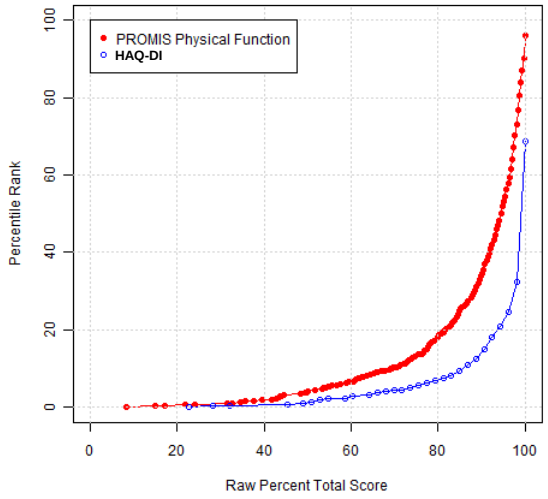

Figure 5.6.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

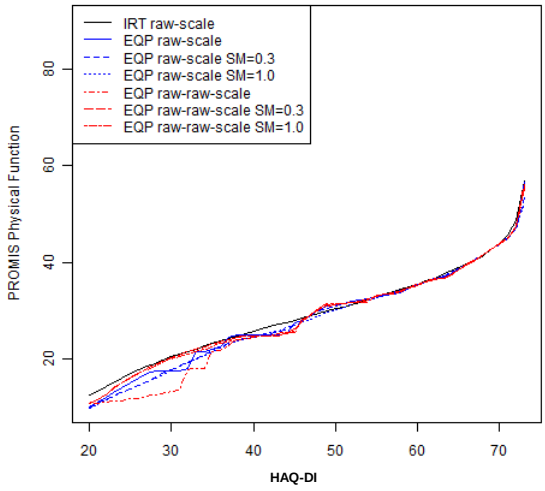

Figure 5.6.10: Equipercentile Linking Functions

5.6.7. 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 5.6.10, 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 5.6.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Physical Function T-scores and HAQ-DI scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.6.10), 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 (0.799), followed by IRT raw-scale (0.802). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.704), followed by IRT raw-scale (5.767).

Table 5.6.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.806	-0.031	5.708	5.704
IRT raw-scale	0.802	-0.045	5.772	5.767
EQP raw-scale SM=0.0	0.802	0.280	5.776	5.778
EQP raw-scale SM=0.3	0.799	2.339	5.882	6.326
EQP raw-scale SM=1.0	0.795	2.994	5.993	6.695
EQP raw-raw-scale SM=0.0	0.801	0.476	5.772	5.787
EQP raw-raw-scale SM=0.3	0.802	0.705	5.764	5.802
EQP raw-raw-scale SM=1.0	0.801	1.209	5.771	5.892

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=622), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=622) over a large number of replications (i.e., 10,000).

Table 5.6.6 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 Physical Function 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, 0.617. That is, the difference between the mean PROMIS Physical Function T-score and the mean equated HAQ-DI T-score based on a similar sample of 75 cases is expected to be around ±1.23 (i.e., 2 x 0.617).

Table 5.6.6: 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	-0.022	1.116	-0.035	0.768	-0.036	0.617
IRT raw-scale	-0.041	1.128	-0.042	0.784	-0.047	0.620
EQP raw-scale SM=0.0	0.275	1.132	0.281	0.777	0.276	0.632
EQP raw-scale SM=0.3	2.335	1.157	2.356	0.794	2.337	0.639
EQP raw-scale SM=1.0	2.971	1.167	3.010	0.815	2.994	0.643
EQP raw-raw-scale SM=0.0	0.463	1.134	0.483	0.777	0.465	0.629
EQP raw-raw-scale SM=0.3	0.716	1.117	0.703	0.774	0.705	0.627
EQP raw-raw-scale SM=1.0	1.207	1.117	1.210	0.770	1.202	0.620

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 16) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.7. PROMIS Physical Function and SF-36/PF

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Physical Function, namely the PROMIS Physical Function (76 items) and SF-36/PF (10 items). PROMIS Physical Function was scaled such that higher scores represent higher levels of Physical Function. 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.

5.7.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 380 for PROMIS Physical Function and 30 for SF-36/PF. Figure 5.7.1 and Figure 5.7.2 graphically display the raw summed score distributions of the two measures. Figure 5.7.3 shows the distribution for the combined. Figure 5.7.4 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 Physical Function and SF-36/PF was 0.91. The disattenuated (corrected for unreliabilities) correlation between PROMIS Physical Function and SF-36/PF was 0.95. The correlations between the combined score and the measures were 1.00 and 0.93 for PROMIS Physical Function and SF-36/PF, respectively.

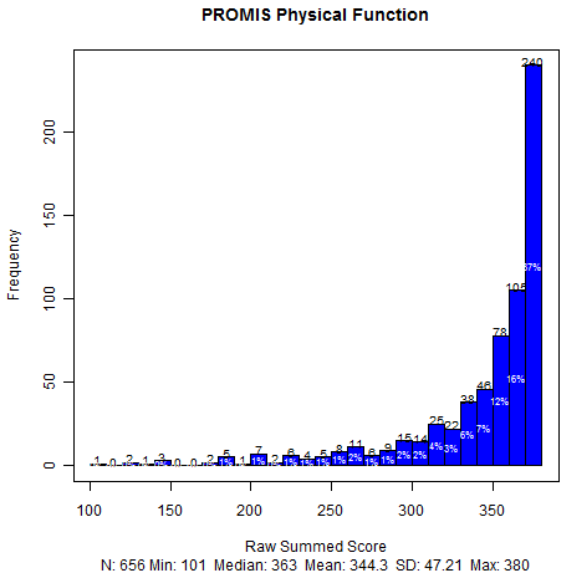

Figure 5.7.1: Raw Summed Score Distribution - PROMIS Instrument

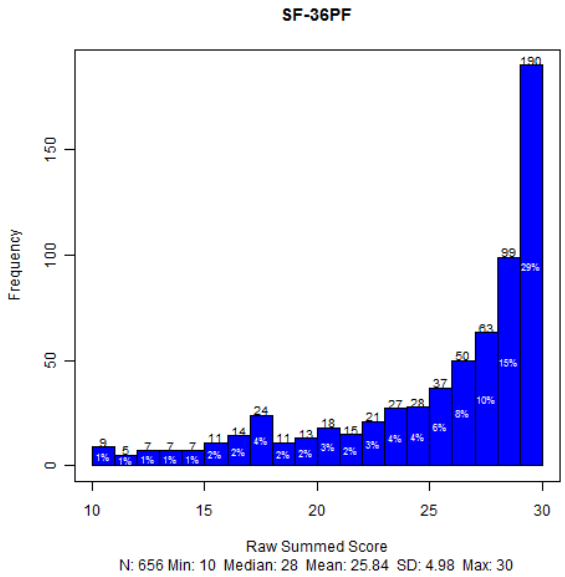

Figure 5.7.2: Raw Summed Score Distribution – Linking Instrument

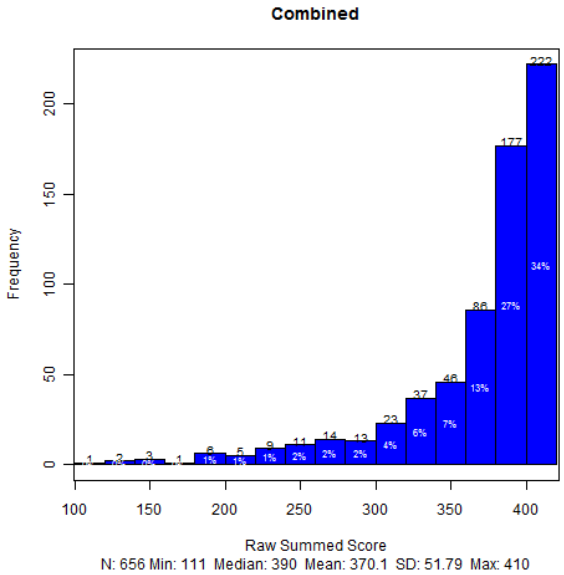

Figure 5.7.3: Raw Summed Score Distribution – Combined

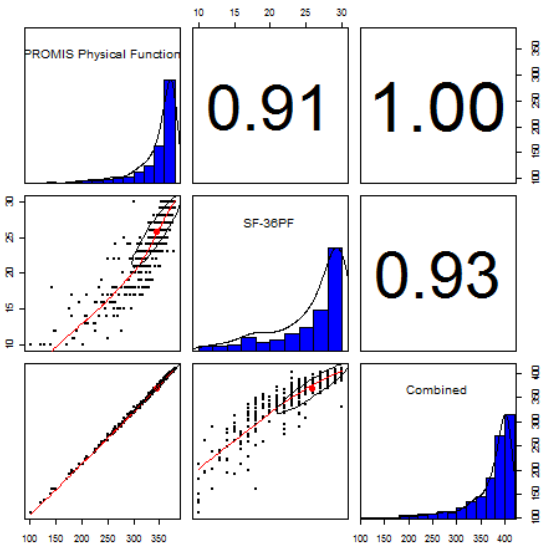

Figure 5.7.4: Scatter Plot Matrix of Raw Summed Scores

5.7.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.7.1 summarizes the results. For PROMIS Physical Function, Cronbach's alpha internal consistency reliability estimate 0.987 and adjusted (corrected for overlap) item-total correlations ranged from 0.517 to 0.868. For SF-36/PF, alpha was 0.929 and adjusted item-total correlations ranged from 0.498 to 0.834. For the 86 items, alpha was 0.988 and adjusted item-total correlations ranged from 0.507 to 0.872.

Table 5.7.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Physical Function	76	0.987	0.517	0.729	0.868
SF-36/PF	10	0.929	0.498	0.736	0.834
Combined	86	0.988	0.507	0.728	0.872

5.7.3. 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 each of the two measures separately and on the combined. Table 5.7.2 summarizes the model fit statistics. For PROMIS Physical Function, the fit statistics were as follows: CFI = 0.977, TLI = 0.976, and RMSEA = 0.043. For SF-36/PF, CFI = 0.99, TLI = 0.987, and RMSEA = 0.112. For the 86 items,

CFI = 0.976, TLI = 0.976, and RMSEA = 0.042. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.7.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Physical Function	76	719	0.977	0.976	0.043
SF-36/PF	10	719	0.990	0.987	0.112
Combined	86	719	0.976	0.976	0.042

5.7.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 86 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 86 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 76 PROMIS Physical Function items served as anchor items to transform the item parameter estimates for the SF-36/PF items onto the PROMIS Physical Function 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 5.7.3 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 items were constrained to their final bank values, while the SF-36/PF items were calibrated under the constraints imposed by the anchor items.

Table 5.7.3: IRT Linking Constants

	A	B
Mean/Mean	1.402	-1.323
Mean/Sigma	1.442	-1.313
Haebara	1.419	-1.310
Stocking-Lord	1.430	-1.307

The item parameter estimates for the SF-36/PF items were linked to the PROMIS Physical Function metric using the transformation constants shown in Table 5.7.3. The SF-36/PF item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Physical Function metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for SF-36/PF as shown in Figure 5.7.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.7.6 displays the differences on the vertical axis.

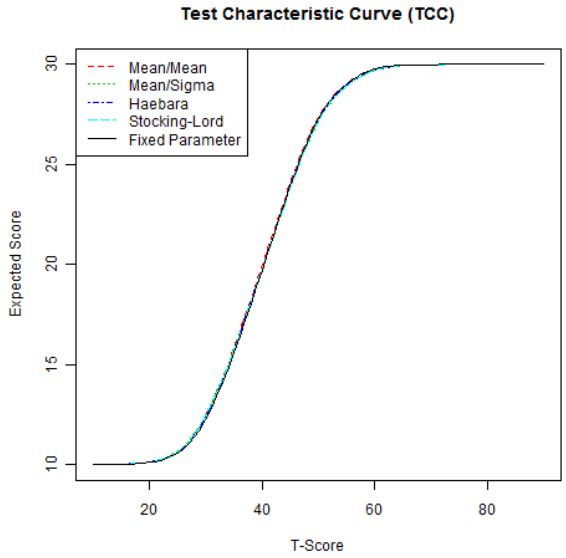

Figure 5.7.5: Test Characteristic Curves (TCC) from Different Linking Methods

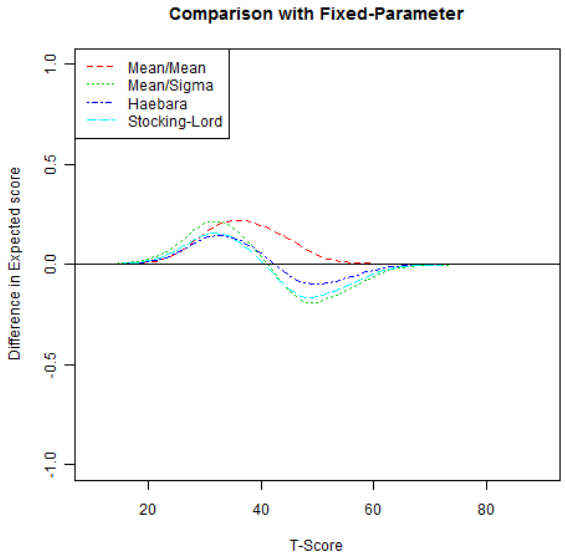

Figure 5.7.6: Difference in Test Characteristic Curves (TCC)

Table 5.7.4 shows the fixed-parameter calibration item parameter estimates for SF-36/PF. The marginal reliability estimate for SF-36/PF based on the item parameter estimates was 0.838. The marginal reliability estimates for PROMIS Physical Function and the combined set were 0.956 and 0.96, respectively. The slope parameter estimates for SF-36/PF ranged from 2.59 to 5.09 with a mean of 3.68. The slope parameter estimates for PROMIS Physical Function ranged from 2.03 to 4.83 with a mean of 3.29. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.7.7 displays the scale information functions for PROMIS Physical Function, SF-36/PF, and the combined set of 86. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.7.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.7.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2
3.745	-0.314	0.529
5.087	-1.363	-0.386
3.884	-1.841	-0.865
3.541	-0.855	-0.063
3.736	-1.748	-0.788
2.588	-1.471	-0.256
3.772	-0.965	-0.148
3.254	-1.464	-0.690
3.500	-1.784	-0.935
3.666	-2.143	-1.582

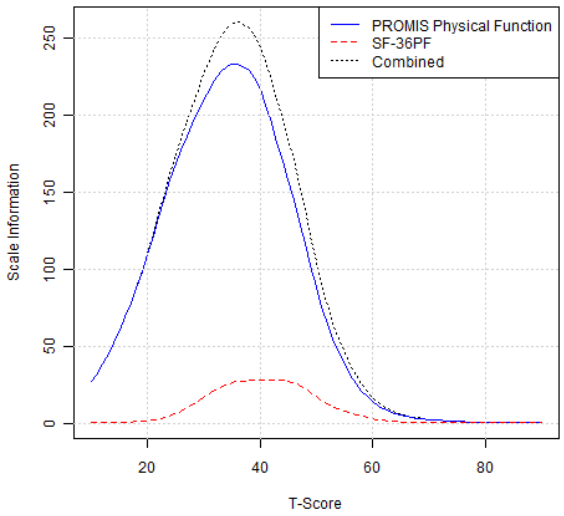

Figure 5.7.7: Comparison of Scale Information Functions

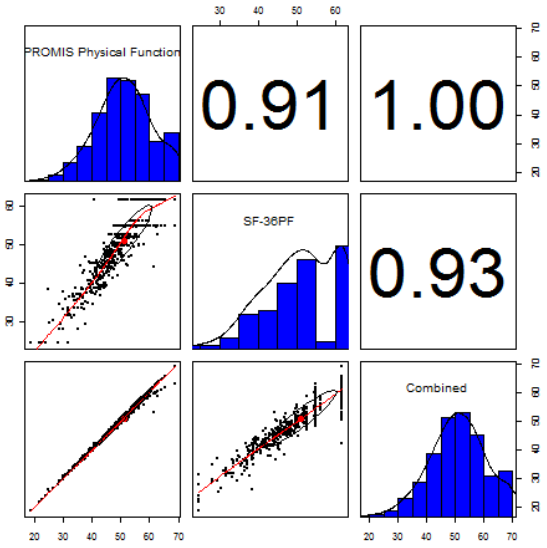

Figure 5.7.8: Comparison of IRT Scaled Scores

5.7.5. 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 SF-36/PF to a scaled score on PROMIS Physical Function can be useful. Based on the SF-36/PF item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 19 can be used to map simple raw summed scores from SF-36/PF to T-score values linked to the PROMIS Physical Function 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.

5.7.6. Equipercentile Linking

We mapped each raw summed score point on SF-36 PF10 to a corresponding scaled score on PROMIS Physical Function by identifying scores on PROMIS Physical Function that have the same percentile ranks as scores on SF-36 PF10. 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 5.2.9 displays the cumulative distribution functions of the measures. Figure 5.2.10 shows the equipercentile linking functions based on raw summed scores, from SF-36/PF to PROMIS Physical Function. 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 20 and Appendix Table 21 show the equipercentile crosswalk tables. The result shown in Appendix Table 20 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 21 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).

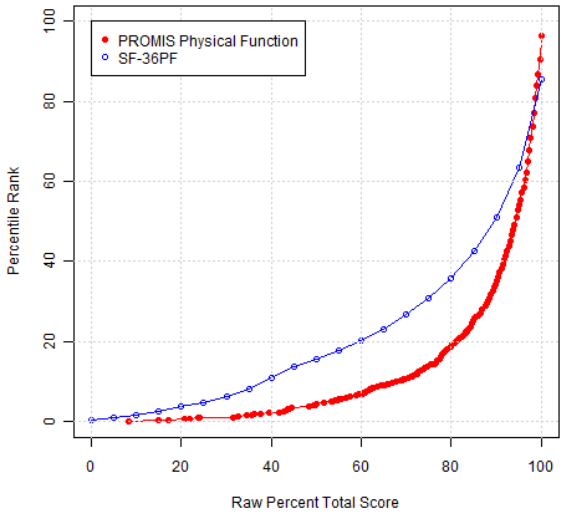

Figure 5.7.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

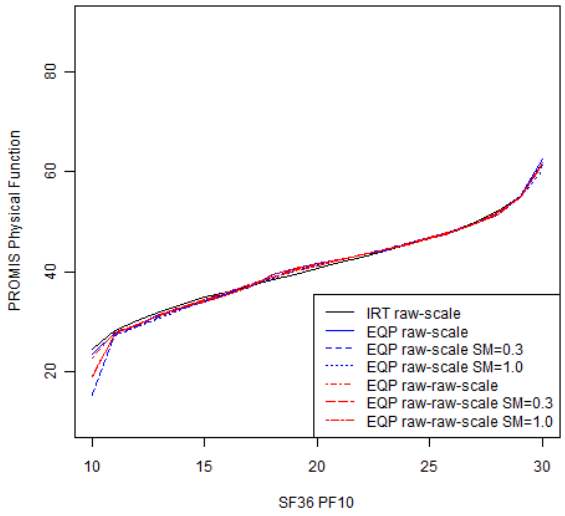

Figure 5.7.10: Equipercentile Linking Functions

5.7.7. 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 5.7.10, 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 5.7.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Physical

Function T-scores and SF-36/PF scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.7.10), 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, EQP raw-raw-scale SM=0.0 produced the best result (0.907), followed by IRT pattern scoring (0.907). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). EQP raw-raw-scale SM=0.0 yielded smallest RMSD (4.109), followed by IRT pattern scoring (4.111).

Table 5.7.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.907	0.032	4.114	4.111
IRT raw-scale	0.906	0.016	4.117	4.114
EQP raw-scale SM=0.0	0.906	-0.154	4.156	4.155
EQP raw-scale SM=0.3	0.904	0.251	4.190	4.195
EQP raw-scale SM=1.0	0.902	0.530	4.196	4.226
EQP raw-raw-scale SM=0.0	0.907	0.052	4.112	4.109
EQP raw-raw-scale SM=0.3	0.906	0.134	4.143	4.142
EQP raw-raw-scale SM=1.0	0.906	0.173	4.139	4.139

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=656), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=656) over a large number of replications (i.e., 10,000).

Table 5.7.6 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 Physical Function 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 raw-scale produced the smallest standard error, 0.446. That is, the difference between the mean PROMIS Physical Function T-score and the mean equated SF-36/PF T-score based on a similar sample of 75 cases is expected to be around ±0.89 (i.e., 2 x 0.446).

Table 5.7.6: 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	0.029	0.808	0.033	0.567	0.029	0.447
IRT raw-scale	0.016	0.804	0.020	0.552	0.012	0.446
EQP raw-scale SM=0.0	-0.152	0.813	-0.153	0.567	-0.154	0.454
EQP raw-scale SM=0.3	0.261	0.829	0.247	0.575	0.252	0.458
EQP raw-scale SM=1.0	0.526	0.829	0.527	0.573	0.539	0.456
EQP raw-raw-scale SM=0.0	0.044	0.797	0.055	0.556	0.054	0.446
EQP raw-raw-scale SM=0.3	0.137	0.815	0.130	0.564	0.134	0.452
EQP raw-raw-scale SM=1.0	0.178	0.811	0.178	0.560	0.175	0.452

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 19) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.8. PROMIS Fatigue and FACIT-F

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Fatigue, namely the PROMIS Fatigue item bank (95 items) and FACIT-F (13 items). PROMIS Fatigue was scaled such that higher scores represent higher levels of Fatigue; for the FACIT-F, higher scores represent lower levels of Fatigue. 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.

5.8.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 475 for PROMIS Fatigue and 53 for FACIT-F. Figure 5.8.1 and Figure 5.8.2 graphically display the raw summed score distributions of the two measures. Figure 5.8.3 shows the distribution for the combined. Figure 5.8.4 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 Fatigue and FACIT-F was -0.96. The disattenuated (corrected for unreliabilities) correlation between PROMIS Fatigue and FACIT-F was -0.99. The correlations between the combined score and the measures were 1.00 and 0.96 for PROMIS Fatigue and FACIT-F, respectively.

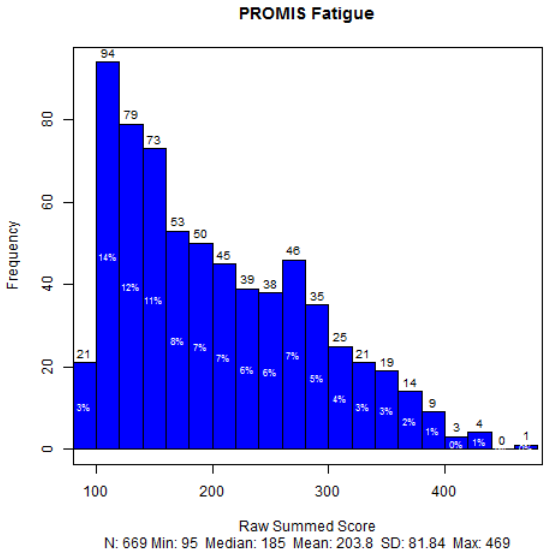

Figure 5.8.1: Raw Summed Score Distribution - PROMIS Instrument

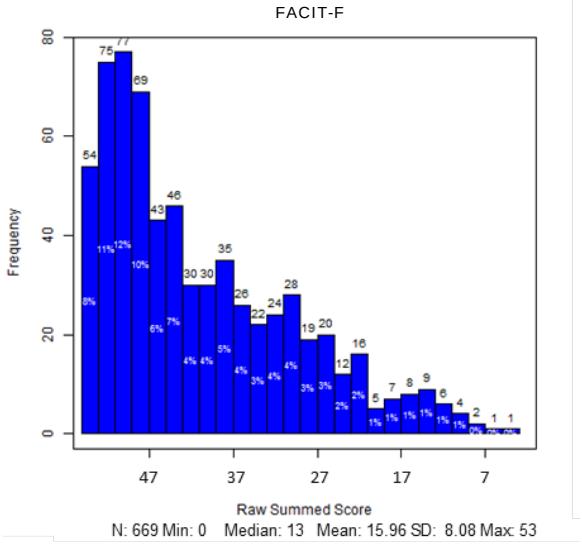

Figure 5.8.2: Raw Summed Score Distribution – Linking Instrument

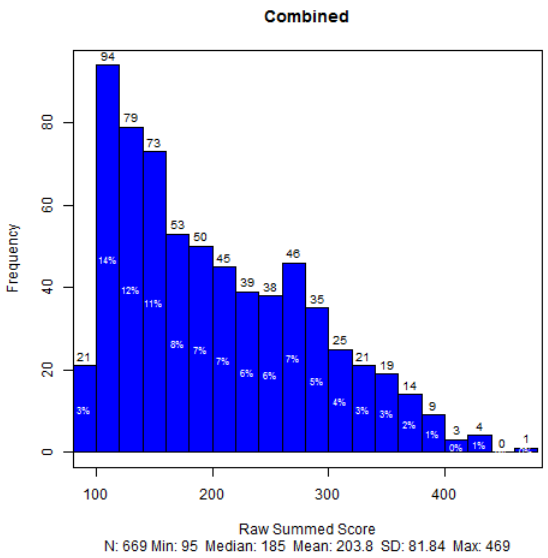

Figure 5.8.3: Raw Summed Score Distribution – Combined

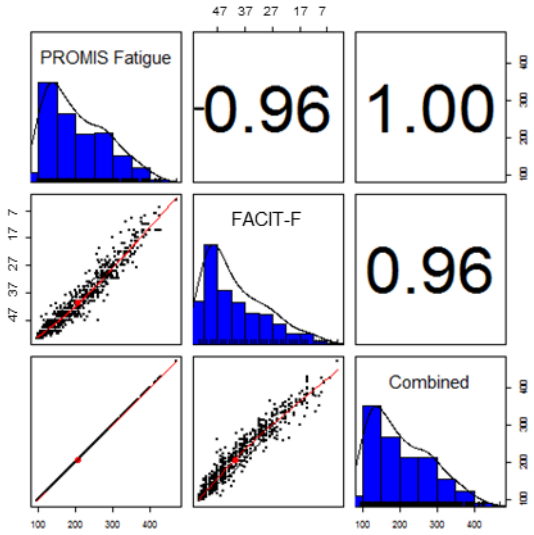

Figure 5.8.4: Scatter Plot Matrix of Raw Summed Scores

5.8.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.8.1 summarizes the results. For PROMIS Fatigue, Cronbach's alpha internal consistency reliability estimate was 0.994 and adjusted (corrected for overlap) item-total correlations ranged from 0.509 to 0.883. For FACIT-F, alpha was 0.958 and adjusted item-total correlations ranged from 0.610 to 0.877. For the 95 items, alpha was 0.994 and adjusted item-total correlations ranged from 0.509 to 0.883.

Table 5.8.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Fatigue	95	0.994	0.509	0.805	0.883
FACIT-F	13	0.958	0.610	0.781	0.877
Combined	95	0.994	0.509	0.805	0.883

5.8.3. 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 each of the two measures separately and on the combined. Table 5.8.2 summarizes the model fit statistics. For PROMIS Fatigue, the fit statistics were as follows: CFI = 0.969, TLI = 0.968, and RMSEA = 0.062. For FACIT-F, CFI = 0.989, TLI = 0.986, and RMSEA = 0.102. For the 95 items, CFI = 0.969, TLI = 0.968, and RMSEA = 0.062. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.8.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Fatigue	95	738	0.969	0.968	0.062
FACIT-F	13	738	0.989	0.986	0.102
Combined	95	738	0.969	0.968	0.062

5.8.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 95 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 95 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 95 PROMIS Fatigue items served as anchor items to transform the item parameter estimates for the FACIT-F items onto the PROMIS Fatigue 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 5.8.3 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 items were constrained to their final bank values, while the FACIT-F items were calibrated under the constraints imposed by the anchor items.

Table 5.8.3: IRT Linking Constants

	A	B
Mean/Mean	1.012	0.588
Mean/Sigma	1.002	0.591
Haebara	1.004	0.595
Stocking-Lord	1.002	0.596

The item parameter estimates for the FACIT-F items were linked to the PROMIS Fatigue metric using the transformation constants shown in Table 5.8.3. The FACIT-F item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Fatigue metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for FACIT-F as shown in Figure 5.8.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.8.6 displays the differences on the vertical axis.

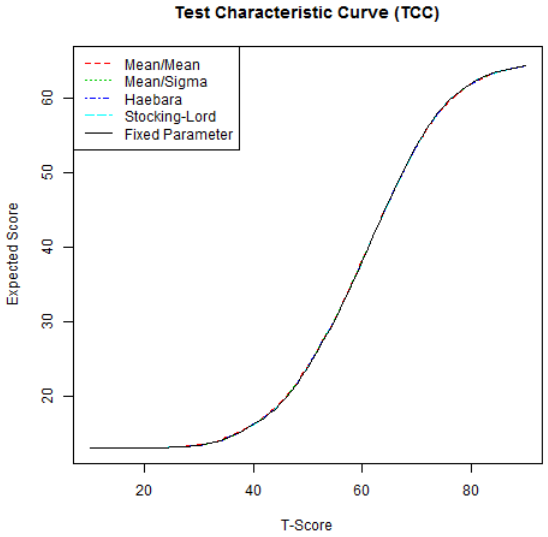

Figure 5.8.5: Test Characteristic Curves (TCC) from Different Linking Methods

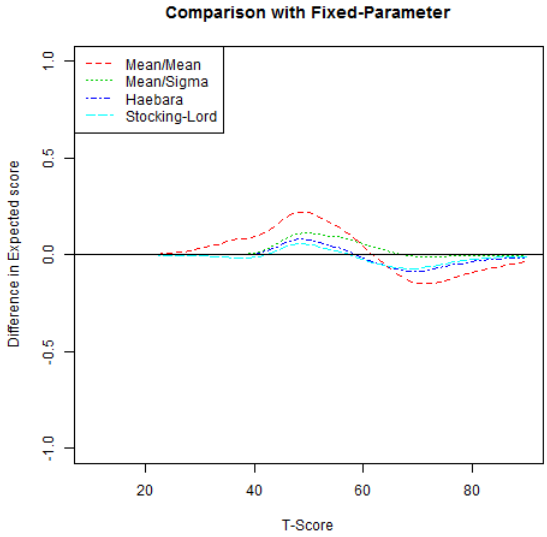

Figure 5.8.6: Difference in Test Characteristic Curves (TCC)

Table 5.8.4 shows the fixed-parameter calibration item parameter estimates for FACIT-F. The marginal reliability estimate for FACIT-F based on the item parameter estimates was 0.945. The marginal reliability estimates for PROMIS Fatigue and the combined set were 0.991 and 0.991, respectively. The slope parameter estimates for FACIT-F ranged from 1.64 to 4.35 with a mean of 3.1. The slope parameter estimates for PROMIS Fatigue ranged from 1.17 to 4.77 with a mean of 3.17. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.8.7 displays the scale information functions for PROMIS Fatigue, FACIT-F, and the combined set of 95. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.8.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.8.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
4.320	-1.140	0.081	0.956	1.770
2.690	0.161	0.916	1.660	2.420
3.270	-0.203	0.744	1.370	2.270
3.300	-1.360	0.099	0.844	1.800
4.350	-0.311	0.504	1.180	1.960
3.400	-0.341	0.582	1.280	2.140
1.640	-0.433	0.874	1.560	2.520
2.310	0.948	1.720	2.470	3.540
2.310	0.753	1.520	2.200	3.020
3.900	-0.088	0.613	1.070	1.430
3.610	0.147	0.829	1.420	1.870
2.710	-1.390	0.003	1.010	2.170
2.550	-0.498	0.597	1.680	2.600

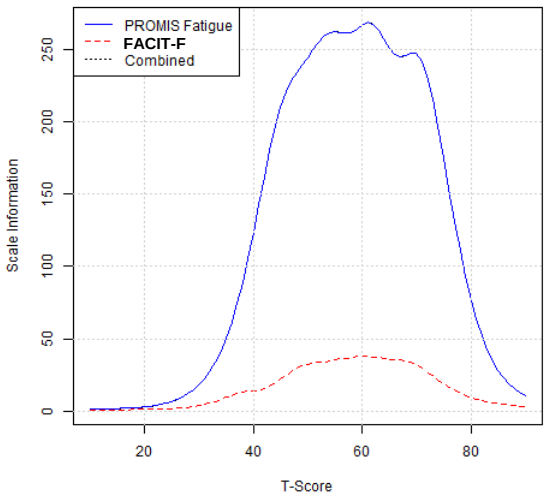

Figure 5.8.7: Comparison of Scale Information Functions

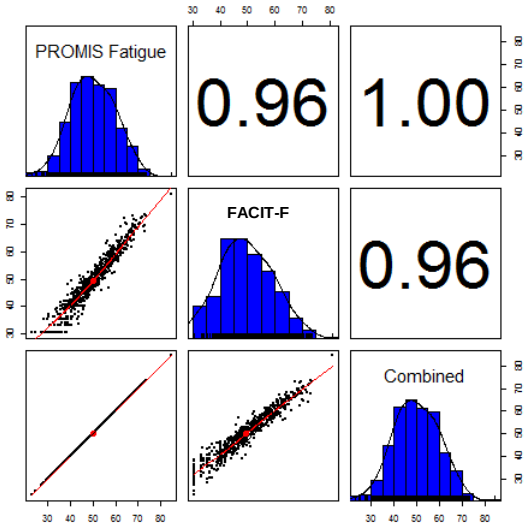

Figure 5.8.8: Comparison of IRT Scaled Scores

5.8.5. 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 FACIT-F to a scaled score on PROMIS Fatigue can be useful. Based on the FACIT-F item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 22 can be used to map simple raw summed scores from FACIT-F to T-score values linked to the PROMIS Fatigue 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.

5.8.6. Equipercentile Linking

We mapped each raw summed score point on FACIT-F to a corresponding scaled score on PROMIS Fatigue by identifying scores on PROMIS Fatigue that have the same percentile ranks as scores on FACIT-F. 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 5.8.9 displays the cumulative distribution functions of the measures. Figure 5.8.10 shows the equipercentile linking functions based on raw summed scores, from FACIT-F to PROMIS Fatigue. 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 23 and Appendix Table 24 show the equipercentile crosswalk tables. The result shown in Appendix Table 23 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 24 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).

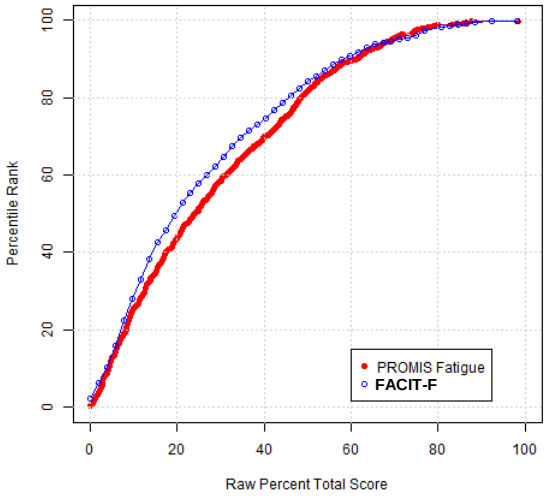

Figure 5.8.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

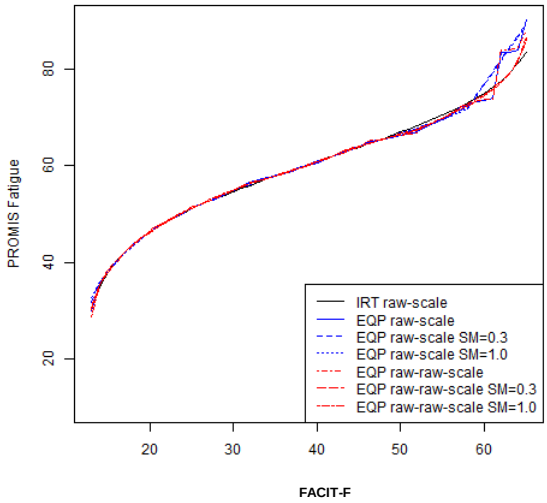

Figure 5.8.10: Equipercentile Linking Functions

5.8.7. 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 5.8.10, 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 5.8.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Fatigue T-scores and FACIT-F scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.8.10), 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 (0.958), followed by EQP raw-scale SM=0.0 (0.955). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (2.867), followed by EQP raw-scale SM=0.3 (2.942).

Table 5.8.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.958	0.472	2.830	2.867
IRT raw-scale	0.955	-0.026	2.952	2.950
EQP raw-scale SM=0.0	0.955	-0.029	2.948	2.946
EQP raw-scale SM=0.3	0.955	-0.127	2.941	2.942
EQP raw-scale SM=1.0	0.954	-0.158	2.944	2.946
EQP raw-raw-scale SM=0.0	0.955	-0.056	2.944	2.943
EQP raw-raw-scale SM=0.3	0.955	-0.054	2.951	2.950
EQP raw-raw-scale SM=1.0	0.954	-0.050	2.978	2.976

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=669), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=669) over a large number of replications (i.e., 10,000).

Table 5.8.6 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 Fatigue 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, 0.31. That is, the difference between the mean PROMIS Fatigue T-score and the mean equated FACIT-F T-score based on a similar sample of 75 cases is expected to be around ±0.62 (i.e., 2 x 0.31).

Table 5.8.6: 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	0.481	0.556	0.471	0.383	0.470	0.310
IRT raw-scale	-0.030	0.577	-0.023	0.402	-0.024	0.323
EQP raw-scale SM=0.0	-0.018	0.584	-0.035	0.399	-0.031	0.322
EQP raw-scale SM=0.3	-0.127	0.576	-0.130	0.401	-0.126	0.319
EQP raw-scale SM=1.0	-0.165	0.581	-0.162	0.400	-0.163	0.315
EQP raw-raw-scale SM=0.0	-0.067	0.585	-0.075	0.399	-0.062	0.319
EQP raw-raw-scale SM=0.3	-0.053	0.580	-0.057	0.403	-0.052	0.322
EQP raw-raw-scale SM=1.0	-0.049	0.578	-0.049	0.407	-0.047	0.325

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 22) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.9. PROMIS Fatigue and SF-36/VT

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Fatigue, namely the PROMIS Fatigue (82 items) and SF-36/VT (4 items). PROMIS Fatigue was scaled such that higher scores represent higher levels of Fatigue; for the SF-36/VT, higher scores represent lower levels of Fatigue. 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.

5.9.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 410 for PROMIS Fatigue and 20 for SF-36/VT. Figure 5.9.1 and Figure 5.9.2 graphically display the raw summed score distributions of the two measures. Figure 5.9.3 shows the distribution for the combined. Figure 5.9.4 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 Fatigue and SF-36/VT was -0.89. The disattenuated (corrected for unreliabilities) correlation between PROMIS Fatigue and SF-36/VT was -0.93. The correlations between the combined score and the measures were 1.00 and 0.90 for PROMIS Fatigue and SF-36/VT, respectively.

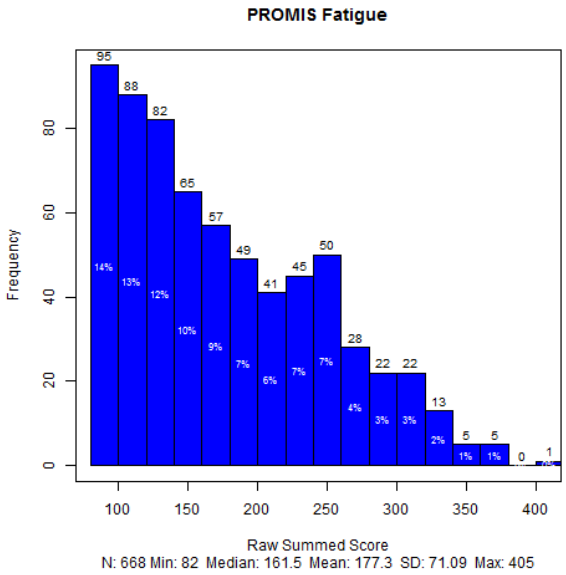

Figure 5.9.1: Raw Summed Score Distribution - PROMIS Instrument

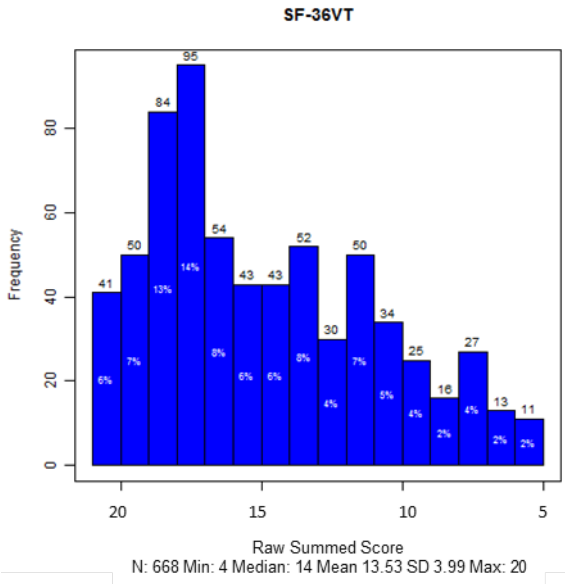

Figure 5.9.2: Raw Summed Score Distribution – Linking Instrument

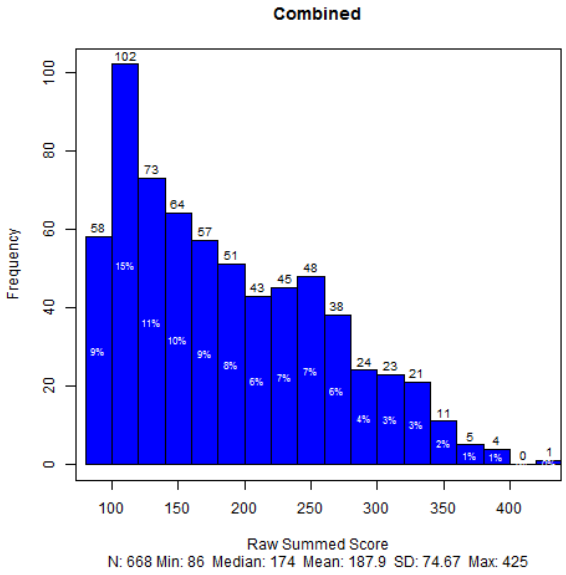

Figure 5.9.3: Raw Summed Score Distribution – Combined

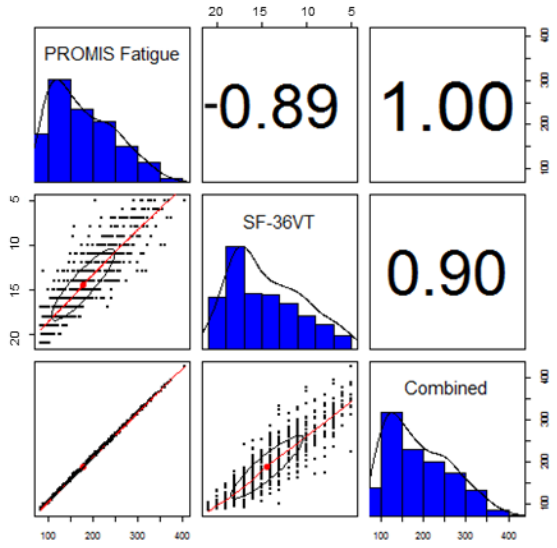

Figure 5.9.4: Scatter Plot Matrix of Raw Summed Scores

5.9.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.9.1 summarizes the results. For PROMIS Fatigue, Cronbach's alpha internal consistency reliability estimate was 0.994 and adjusted (corrected for overlap) item-total correlations ranged from 0.508 to 0.881. For SF-36/VT, alpha was 0.921 and adjusted item-total correlations ranged from 0.806 to 0.828. For the 86 items, alpha was 0.994 and adjusted item-total correlations ranged from 0.511 to 0.883.

Table 5.9.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Fatigue	82	0.994	0.508	0.810	0.881
SF-36/VT	4	0.921	0.806	0.820	0.828
Combined	86	0.994	0.511	0.809	0.883

5.9.3. 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 each of the two measures separately and on the combined. Table 5.9.2 summarizes the model fit statistics. For PROMIS Fatigue, the fit statistics were as follows: CFI = 0.969, TLI = 0.969, and RMSEA = 0.068. For SF-36/VT, CFI = 0.988, TLI = 0.964, and RMSEA = 0.429. For the 86 items, CFI =

0.968, TLI = 0.967, and RMSEA = 0.068. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.9.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Fatigue	82	735	0.969	0.969	0.068
SF-36/VT	4	735	0.988	0.964	0.429
Combined	86	735	0.968	0.967	0.068

5.9.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 86 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 86 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 82 PROMIS Fatigue items served as anchor items to transform the item parameter estimates for the SF-36/VT items onto the PROMIS Fatigue 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 5.9.3 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 items were constrained to their final bank values, while the SF-36/VT items were calibrated under the constraints imposed by the anchor items.

Table 5.9.3: IRT Linking Constants

	A	B
Mean/Mean	1.014	0.529
Mean/Sigma	1.002	0.533
Haebara	1.006	0.536
Stocking-Lord	1.002	0.539

The item parameter estimates for the SF-36/VT items were linked to the PROMIS Fatigue metric using the transformation constants shown in Table 5.9.3. The SF-36/VT item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Fatigue metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for SF-36/VT as shown in Figure 5.9.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.9.6 displays the differences on the vertical axis.

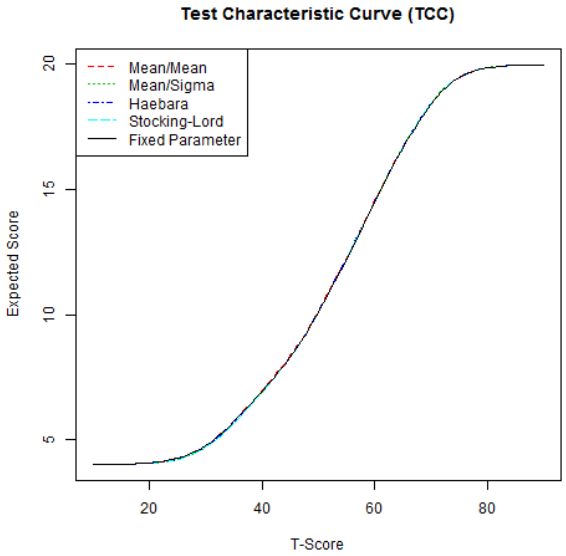

Figure 5.9.5: Test Characteristic Curves (TCC) from Different Linking Methods

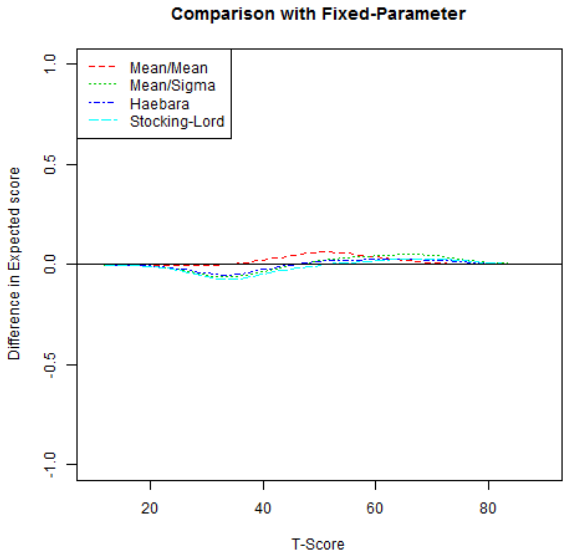

Figure 5.9.6: Difference in Test Characteristic Curves (TCC)

Table 5.9.4 shows the fixed-parameter calibration item parameter estimates for SF-36/VT. The marginal reliability estimate for SF-36/VT based on the item parameter estimates was 0.878. The marginal reliability estimates for PROMIS Fatigue and the combined set were 0.989 and 0.991, respectively. The slope parameter estimates for SF-36/VT ranged from 2.28 to 3.23 with a mean of 2.83. The slope parameter estimates for PROMIS Fatigue ranged from 1.17 to 4.77 with a mean of 3.19. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.9.7 displays the scale information functions for PROMIS Fatigue, SF-36/VT, and the combined set of 86. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.9.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.9.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
2.281	-1.367	0.176	0.897	1.661
2.736	-1.801	-0.154	0.669	1.431
3.234	-0.733	0.204	1.102	2.057
3.061	-1.467	0.036	0.889	1.820

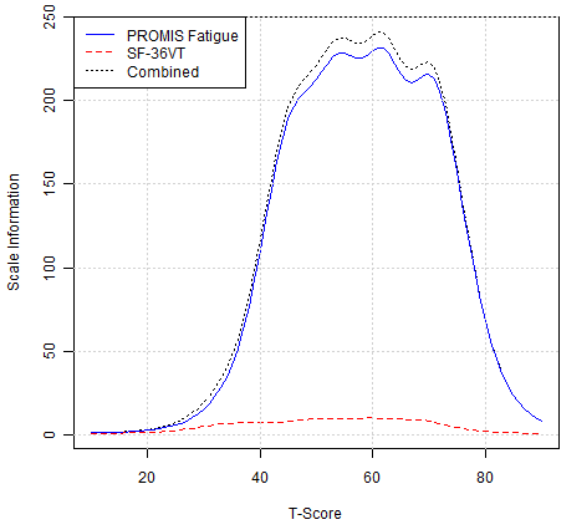

Figure 5.9.7: Comparison of Scale Information Functions

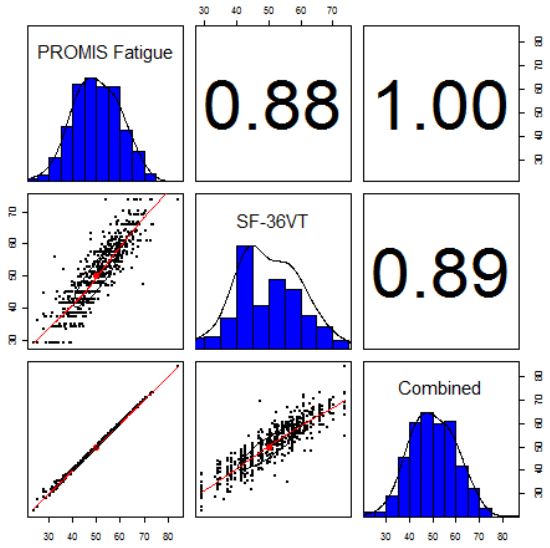

Figure 5.9.8: Comparison of IRT Scaled Scores

5.9.5. 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 SF-36/VT to a scaled score on PROMIS Fatigue can be useful. Based on the SF-36/VT item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 25 can be used to map simple raw summed scores from SF-36/VT to T-score values linked to the PROMIS Fatigue 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.

5.9.6. Equipercentile Linking

We mapped each raw summed score point on SF-36/VT to a corresponding scaled score on PROMIS Fatigue by identifying scores on PROMIS Fatigue that have the same percentile ranks as scores on SF-36/VT. 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 5.9.9 displays the cumulative distribution functions of the measures. Figure 5.9.10 shows the equipercentile linking functions based on raw summed scores from SF-36/VT to PROMIS Fatigue. 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 26 and Appendix Table 27 show the equipercentile crosswalk tables. The result shown in Appendix Table 26 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 27 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).

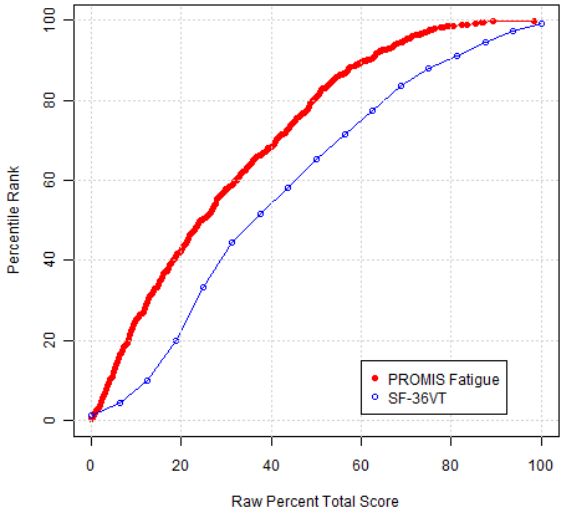

Figure 5.9.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

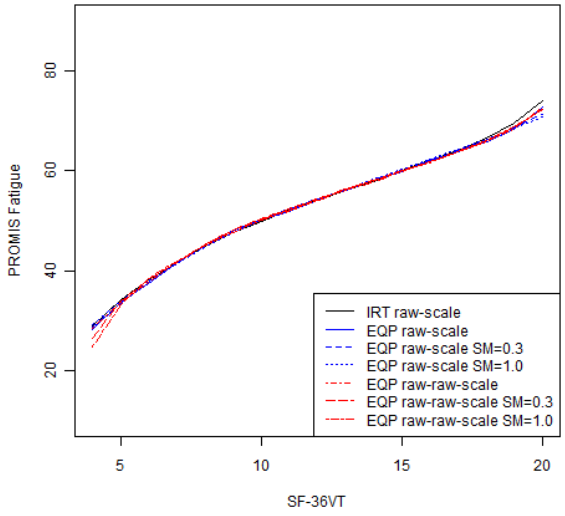

Figure 5.9.10: Equipercentile Linking Functions

5.9.7. 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 5.9.10, 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 5.9.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Fatigue T-

scores and SF-36/VT scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.9.10), 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 (0.881), followed by EQP raw-scale SM=1.0 (0.877). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (4.804), followed by EQP raw-scale SM=1.0 (4.855).

Table 5.9.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.881	-0.071	4.807	4.804
IRT raw-scale	0.876	-0.025	4.895	4.891
EQP raw-scale SM=0.0	0.876	0.024	4.877	4.874
EQP raw-scale SM=0.3	0.877	0.020	4.861	4.858
EQP raw-scale SM=1.0	0.877	-0.004	4.859	4.855
EQP raw-raw-scale SM=0.0	0.876	-0.013	4.867	4.863
EQP raw-raw-scale SM=0.3	0.876	-0.002	4.898	4.894
EQP raw-raw-scale SM=1.0	0.874	0.111	4.971	4.968

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=668), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=668) over a large number of replications (i.e., 10,000).

Table 5.9.6 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 Fatigue 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, 0.52. That is, the difference between the mean PROMIS Fatigue T-score and the mean equated SF-36/VT T-score based on a similar sample of 75 cases is expected to be around ±1.04 (i.e., 2 x 0.52).

Table 5.9.6: 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	-0.056	0.952	-0.081	0.655	-0.068	0.520
IRT raw-scale	-0.035	0.957	-0.025	0.661	-0.032	0.537
EQP raw-scale SM=0.0	0.019	0.961	0.018	0.668	0.026	0.524
EQP raw-scale SM=0.3	0.032	0.950	0.023	0.659	0.025	0.529
EQP raw-scale SM=1.0	0.003	0.943	-0.011	0.661	-0.007	0.534
EQP raw-raw-scale SM=0.0	-0.024	0.959	-0.021	0.660	-0.018	0.532
EQP raw-raw-scale SM=0.3	0.004	0.960	0.004	0.673	-0.002	0.532
EQP raw-raw-scale SM=1.0	0.130	0.983	0.109	0.681	0.104	0.548

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 25) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.10. PROMIS Pain and BPI Severity

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Pain, namely the PROMIS Pain Interference (40 items) and BPI Severity (4 items). PROMIS Pain was scaled such that higher scores represent higher levels of Pain. 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 200 for PROMIS Pain and 16 for BPI Severity. 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 Pain and BPI Severity was 0.77. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pain and BPI Severity was 0.83. The correlations between the combined score and the measures were 1.00 and 0.80 for PROMIS Pain and BPI Severity, respectively.

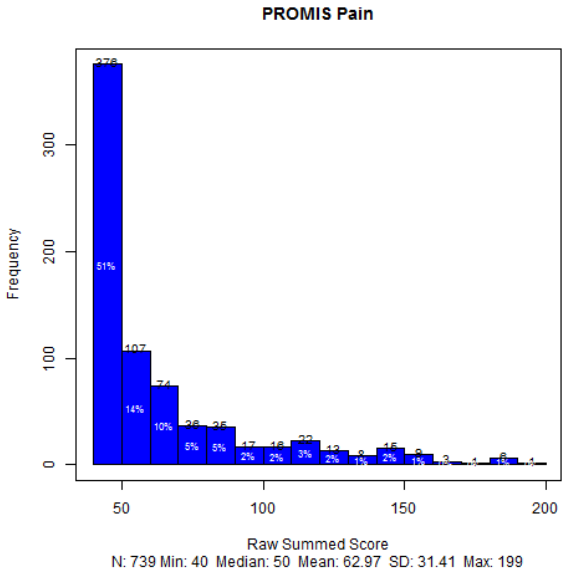

Figure 5.10.1: Raw Summed Score Distribution - PROMIS Instrument

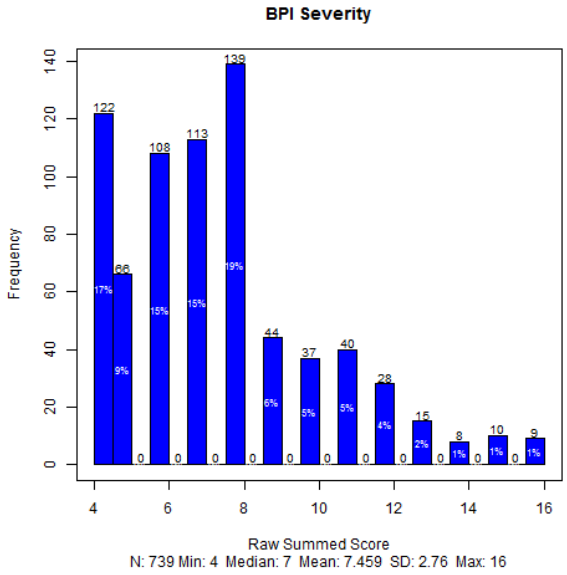

Figure 5.10.2: Raw Summed Score Distribution – Linking Instrument

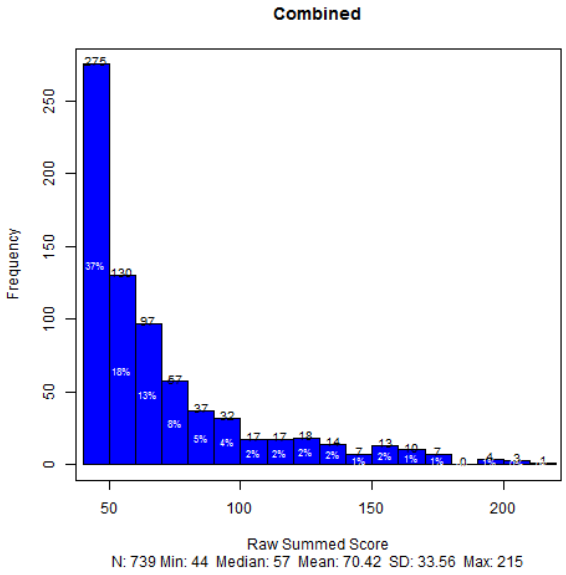

Figure 5.10.3: Raw Summed Score Distribution – Combined

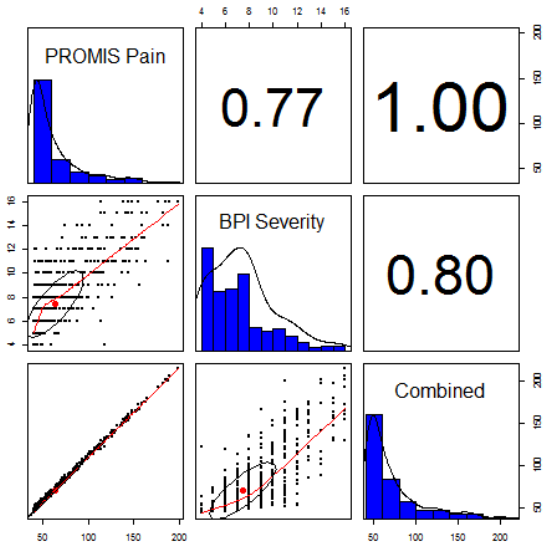

Figure 5.10.4: Scatter Plot Matrix of Raw Summed Scores

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 Pain, Cronbach's alpha internal consistency reliability estimate was 0.987 and adjusted (corrected for overlap) item-total correlations ranged from 0.611 to 0.896. For BPI Severity, alpha was 0.867 and adjusted item-total correlations ranged from 0.670 to 0.807. For the 44 items, alpha was 0.987 and adjusted item-total correlations ranged from 0.573 to 0.893.

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 Pain	40	0.987	0.611	0.809	0.896
BPI Severity	4	0.867	0.670	0.734	0.807
Combined	44	0.987	0.573	0.796	0.893

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 each of the two measures separately and on the combined. Table 5.10.2 summarizes the model fit statistics. For PROMIS Pain, the fit statistics were as follows: CFI = 0.975, TLI = 0.974, and RMSEA = 0.085. For BPI Severity, CFI = 0.998, TLI = 0.994, and RMSEA = 0.107. For the 44 items, CFI = 0.972, TLI = 0.97, and RMSEA = 0.082. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.10.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Pain	40	780	0.975	0.974	0.085
BPI SEVERITY	4	780	0.998	0.994	0.107
Combined	44	780	0.972	0.970	0.082

5.10.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 44 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 44 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 40 PROMIS Pain items served as anchor items to transform the item parameter estimates for the BPI Severity items onto the PROMIS Pain 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 items were constrained to their final bank values, while the BPI Severity items were calibrated under the constraints imposed by the anchor items.

Table 5.10.3: IRT Linking Constants

	A	B
Mean/Mean	1.282	0.805
Mean/Sigma	1.297	0.799
Haebara	1.296	0.810
Stocking-Lord	1.291	0.803

The item parameter estimates for the BPI Severity items were linked to the PROMIS Pain metric using the transformation constants shown in Table 5.10.3. The BPI Severity item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Pain metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for BPI Severity 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.

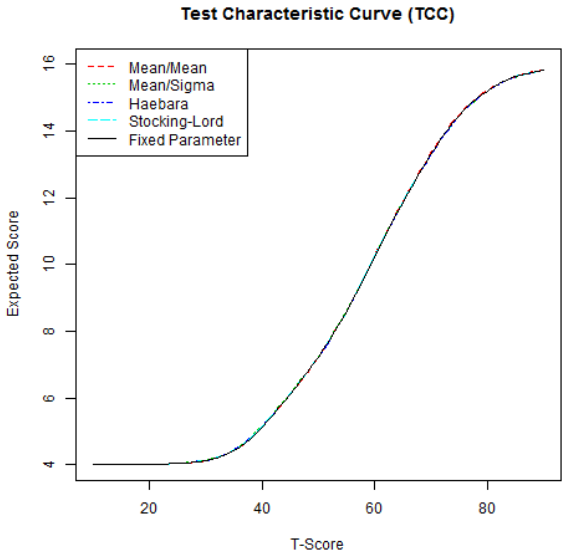

Figure 5.10.5: Test Characteristic Curves (TCC) from Different Linking Methods

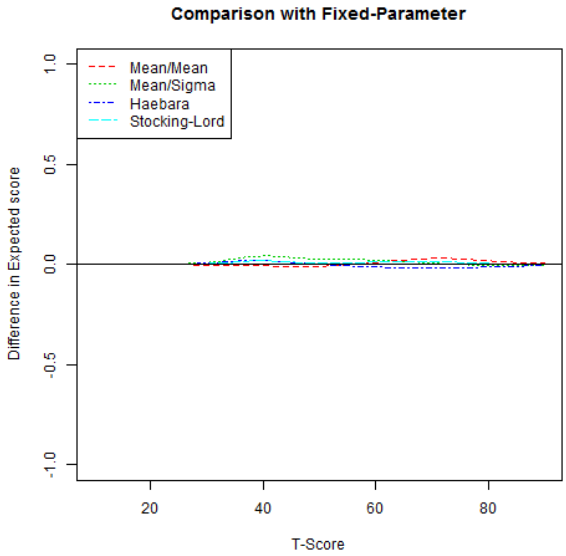

Figure 5.10.6: Difference in Test Characteristic Curves (TCC)

Table 5.10.4 shows the fixed-parameter calibration item parameter estimates for BPI Severity. The marginal reliability estimate for BPI Severity based on the item parameter estimates was 0.811. The marginal reliability estimates for PROMIS Pain and the combined set were 0.883 and 0.937, respectively. The slope parameter estimates for BPI Severity ranged from 1.72 to 2.98 with a mean of 2.45. The slope parameter estimates for PROMIS Pain ranged from 2.2 to 6.53 with a mean of 4.08. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.10.7 displays the scale information functions for PROMIS Pain, BPI Severity, and the combined set of 40. 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

Slope	Threshold 1	Threshold 2	Threshold 3
2.976	-1.069	0.440	0.827
1.722	0.169	2.188	2.758
2.747	-0.755	1.093	1.915
2.345	-0.011	1.303	2.128

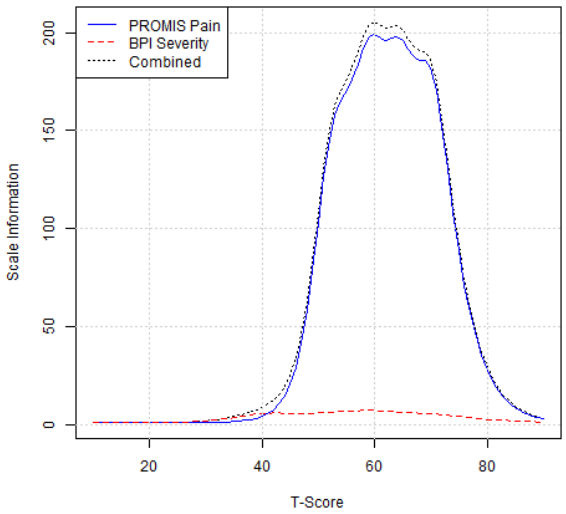

Figure 5.10.7: Comparison of Scale Information Functions

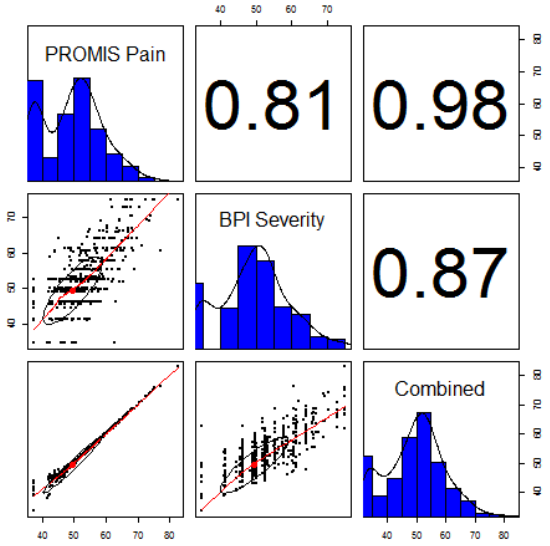

Figure 5.10.8: Comparison of IRT Scaled Scores

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 BPI Severity to a scaled score on PROMIS Pain can be useful. Based on the BPI Severity 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 BPI Severity to T-score values linked to the PROMIS Pain 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 BPI Severity to a corresponding scaled score on PROMIS Pain by identifying scores on PROMIS Pain that have the same percentile ranks as scores on BPI Severity. 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 BPI Severity to PROMIS Pain. 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).

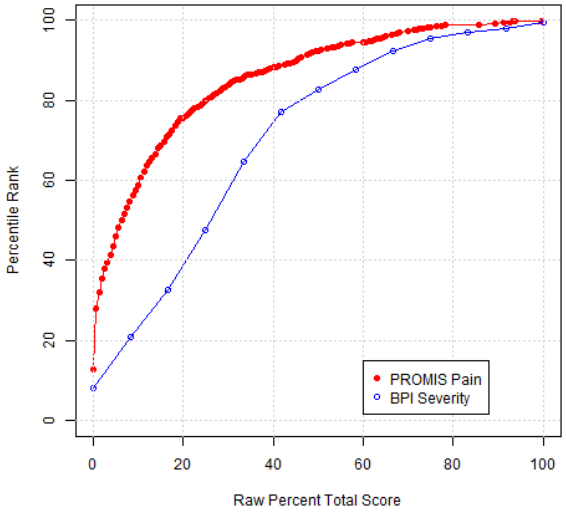

Figure 5.10.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

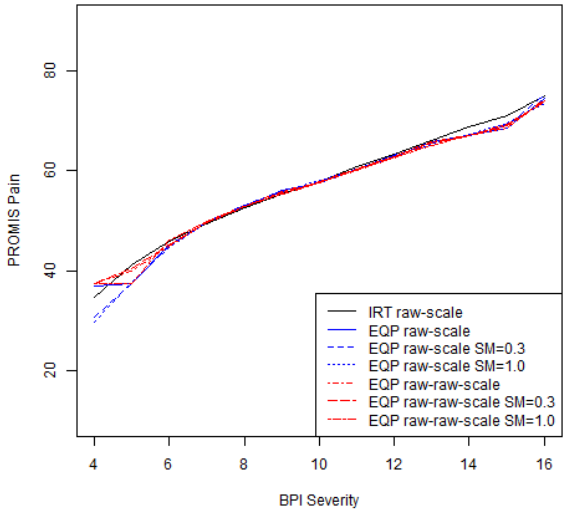

Figure 5.10.10: Equipercentile Linking Functions

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 Pain T-scores

and BPI Severity 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 (0.806), followed by IRT raw-scale (0.802). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). EQP raw-raw-scale SM=0.3 yielded smallest RMSD (5.676), followed by EQP raw-raw-scale SM=0.0 (5.72).

Table 5.10.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.806	0.070	5.829	5.826
IRT raw-scale	0.802	0.183	5.875	5.874
EQP raw-scale SM=0.0	0.796	0.214	5.866	5.866
EQP raw-scale SM=0.3	0.797	1.187	6.522	6.625
EQP raw-scale SM=1.0	0.795	1.405	6.691	6.833
EQP raw-raw-scale SM=0.0	0.798	-0.066	5.723	5.720
EQP raw-raw-scale SM=0.3	0.799	-0.148	5.678	5.676
EQP raw-raw-scale SM=1.0	0.796	0.029	5.776	5.772

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=739), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=739) 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 Pain 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, EQP raw-raw-scale SM=0.0 produced the smallest standard error, 0.625. That is, the difference between the mean PROMIS Pain T-score and the mean equated BPI Severity T-score based on a similar sample of 75 cases is expected to be around ±1.25 (i.e., 2 x 0.625).

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	0.067	1.155	0.072	0.800	0.076	0.640
IRT raw-scale	0.178	1.146	0.179	0.800	0.182	0.646
EQP raw-scale SM=0.0	0.214	1.161	0.207	0.789	0.213	0.640
EQP raw-scale SM=0.3	1.152	1.289	1.187	0.878	1.180	0.711
EQP raw-scale SM=1.0	1.385	1.322	1.402	0.908	1.415	0.732
EQP raw-raw-scale SM=0.0	-0.082	1.133	-0.067	0.782	-0.066	0.625
EQP raw-raw-scale SM=0.3	-0.163	1.122	-0.151	0.777	-0.148	0.626
EQP raw-raw-scale SM=1.0	0.029	1.131	0.035	0.804	0.027	0.636

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.

5.11. PROMIS Pain and BPI Interference

Note: This linking analysis has been revised since its initial publication in Volume 1 (2012). Interested readers may obtain updated results on the Linking Tables and Publications sections of the prosettastone.org website:

http://www.prosettastone.org/LinkingTables1/Pages/default.aspx

http://www.prosettastone.org/PublicationsPresentations/Pages/default.aspx

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Pain, namely the PROMIS Pain Interference (40 items) and BPI Interference (7 items). PROMIS Pain was scaled such that higher scores represent higher levels of Pain. 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.

5.11.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 200 for PROMIS Pain and 28 for BPI Interference. Figure 5.11.1 and Figure 5.11.2 graphically display the raw summed score distributions of the two measures. Figure 5.11.3 shows the distribution for the combined. Figure 5.11.4 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 Pain and BPI Interference was 0.91. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pain and BPI Interference was 0.95. The correlations between the combined score and the measures were 1.00 and 0.93 for PROMIS Pain and BPI Interference, respectively.

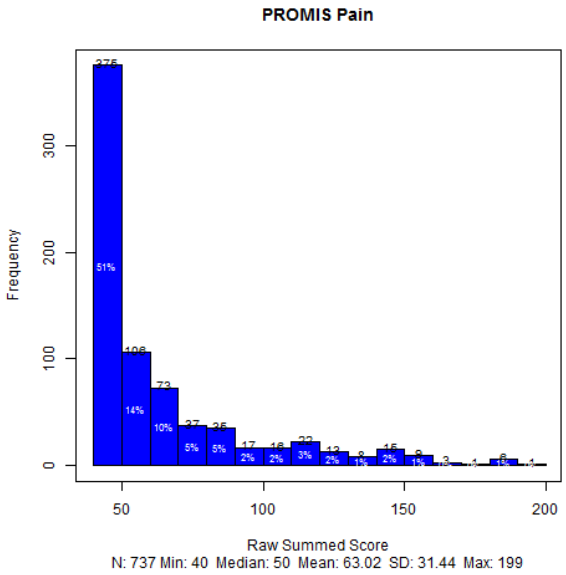

Figure 5.11.1: Raw Summed Score Distribution - PROMIS Instrument

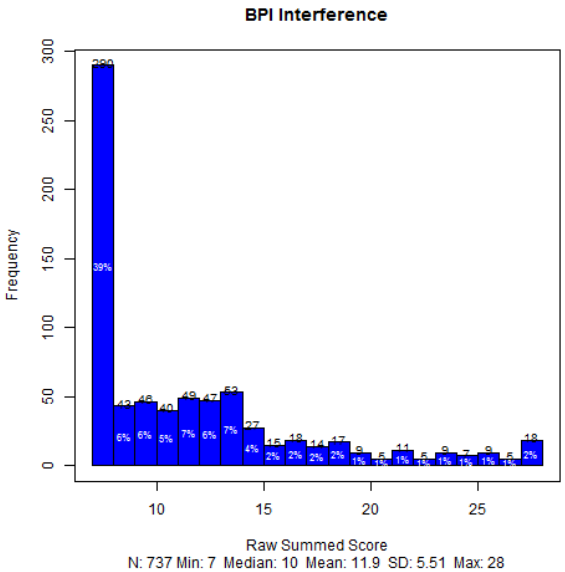

Figure 5.11.2: Raw Summed Score Distribution – Linking Instrument

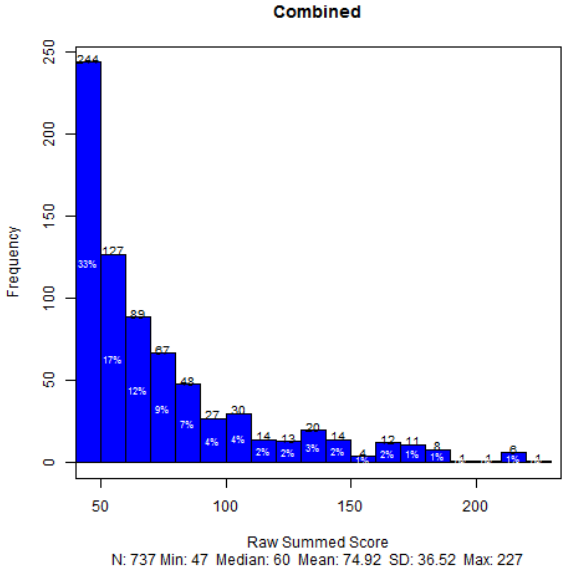

Figure 5.11.3: Raw Summed Score Distribution – Combined

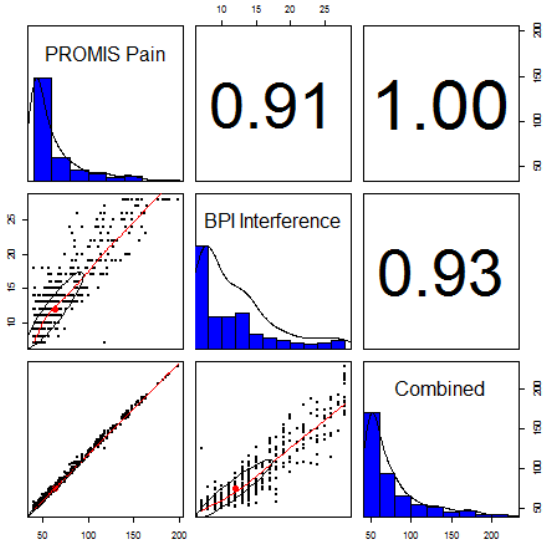

Figure 5.11.4: Scatter Plot Matrix of Raw Summed Scores

5.11.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.11.1 summarizes the results. For PROMIS Pain, Cronbach's alpha internal consistency reliability estimate was 0.987 and adjusted (corrected for overlap) item-total correlations ranged from 0.611 to 0.896. For BPI Interference, alpha was 0.939 and adjusted item-total correlations ranged from 0.708 to 0.867. For the 47 items, alpha was 0.989 and adjusted item-total correlations ranged from 0.609 to 0.895.

Table 5.11.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Pain	40	0.987	0.611	0.809	0.896
BPI Interference	7	0.939	0.708	0.803	0.867
Combined	47	0.989	0.609	0.807	0.895

5.11.3. 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 each of the two measures separately and on the combined. Table 5.11.2 summarizes the model fit statistics. For PROMIS Pain, the fit statistics were as follows: CFI = 0.975, TLI = 0.974, and RMSEA = 0.085. For BPI Interference, CFI = 0.990, TLI = 0.985, and RMSEA = 0.156. For the 47 items,

CFI = 0.971, TLI = 0.97, and RMSEA = 0.082. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.11.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Pain	40	778	0.975	0.974	0.085
BPI Interference	7	778	0.990	0.985	0.156
Combined	47	778	0.971	0.970	0.082

5.11.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 47 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 47 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 40 PROMIS Pain items served as anchor items to transform the item parameter estimates for the BPI Interference items onto the PROMIS Pain 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 5.11.3 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 items were constrained to their final bank values, while the BPI Interference items were calibrated under the constraints imposed by the anchor items.

Table 5.11.3: IRT Linking Constants

	A	B
Mean/Mean	1.210	0.774
Mean/Sigma	1.229	0.766
Haebara	1.227	0.777
Stocking-Lord	1.223	0.770

The item parameter estimates for the BPI Interference items were linked to the PROMIS Pain metric using the transformation constants shown in Table 5.11.3. The BPI Interference item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Pain metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for BPI Interference as shown in Figure 5.11.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.11.6 displays the differences on the vertical axis.

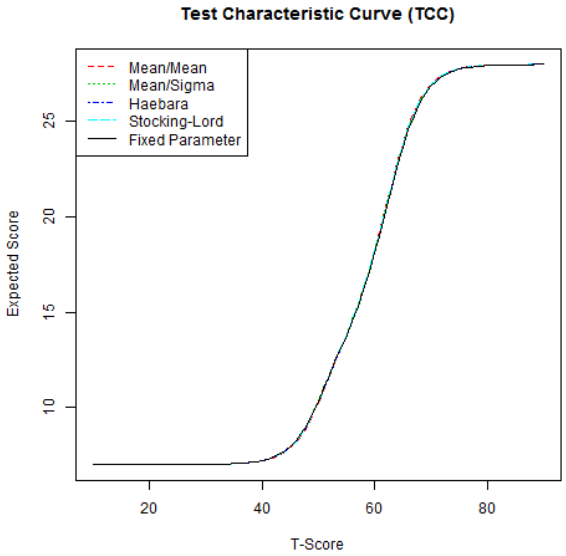

Figure 5.11.5: Test Characteristic Curves (TCC) from Different Linking Methods

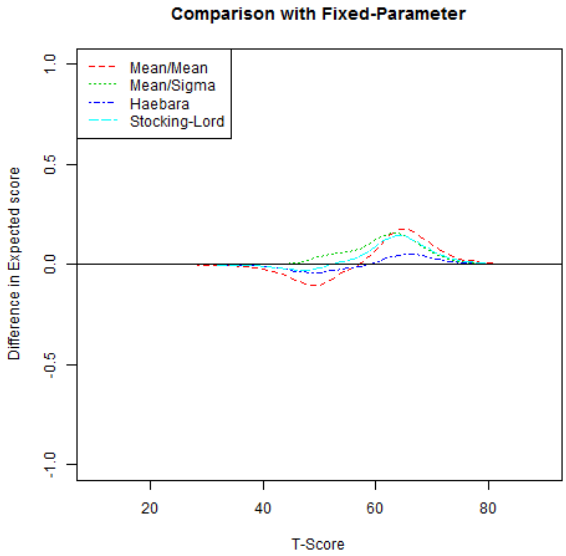

Figure 5.11.6: Difference in Test Characteristic Curves (TCC)

Table 5.11.4 shows the fixed-parameter calibration item parameter estimates for BPI Interference. The marginal reliability estimate for BPI Interference based on the item parameter estimates was 0.827. The marginal reliability estimates for PROMIS Pain and the combined set were 0.883 and 0.905, respectively. The slope parameter estimates for BPI Interference ranged from 2.48 to 4.4 with a mean of 3.65. The slope parameter estimates for PROMIS Pain ranged from 2.20 to 6.53 with a mean of 4.08.We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.11.7 displays the scale information functions for PROMIS Pain, BPI Interference, and the combined set of 47. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.11.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.11.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3
4.309	0.038	1.036	1.437
3.137	-0.086	1.068	1.499
2.835	0.111	0.953	1.303
4.371	0.032	0.951	1.322
3.984	0.403	1.316	1.642
2.481	0.001	1.078	1.371
4.404	0.059	1.060	1.349

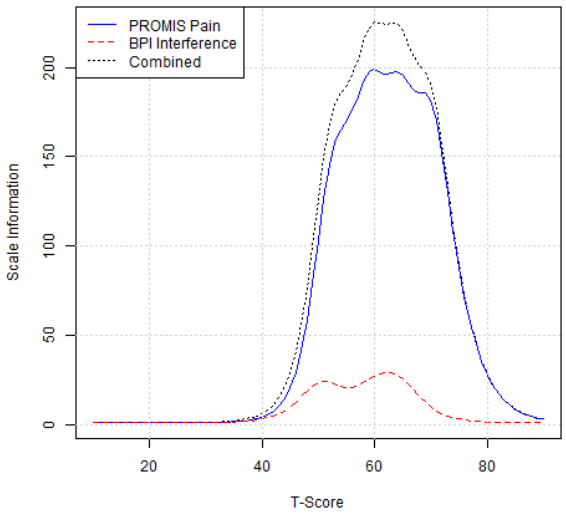

Figure 5.11.7: Comparison of Scale Information Functions

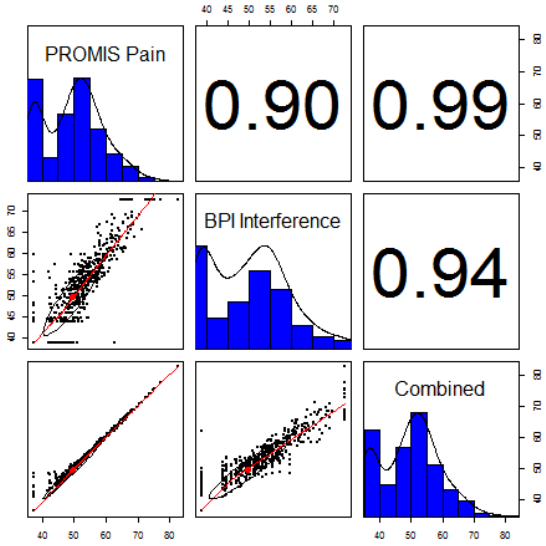

Figure 5.11.8: Comparison of IRT Scaled Scores

5.11.5. 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 BPI Interference to a scaled score on PROMIS Pain can be useful. Based on the BPI Interference item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 31 can be used to map simple raw summed scores from BPI Interference to T-score values linked to the PROMIS Pain 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.

5.11.6. Equipercentile Linking

We mapped each raw summed score point on BPI Interference to a corresponding scaled score on PROMIS Pain by identifying scores on PROMIS Pain that have the same percentile ranks as scores on BPI Interference. 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 5.11.9 displays the cumulative distribution functions of the measures. Figure 5.11.10 shows the equipercentile linking functions based on raw summed scores, from BPI Interference to PROMIS Pain. 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 32 and Appendix Table 33 show the equipercentile crosswalk tables. The result shown in Appendix Table 32 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 33 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).

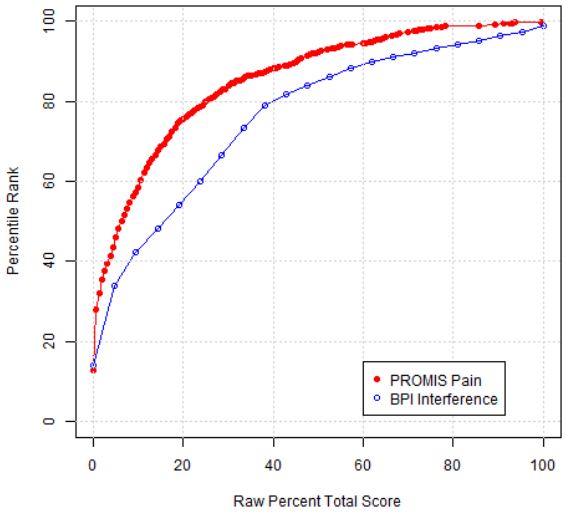

Figure 5.11.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

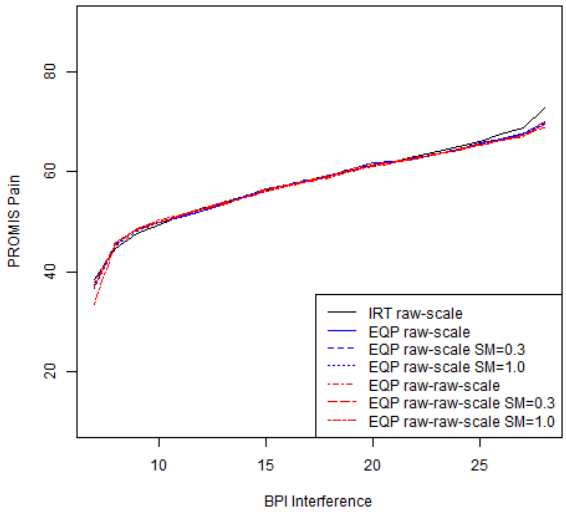

Figure 5.11.10: Equipercentile Linking Functions

5.11.7. 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 5.11.10, 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 5.11.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Pain T-scores

and BPI Interference scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.11.10), 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 raw-scale produced the best result (0.903), followed by IRT pattern scoring (0.902). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT raw-scale yielded smallest RMSD (4.03), followed by EQP raw-raw-scale SM=0.0 (4.056).

Table 5.11.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.902	-0.065	4.060	4.057
IRT raw-scale	0.903	-0.015	4.032	4.030
EQP raw-scale SM=0.0	0.900	0.389	4.142	4.157
EQP raw-scale SM=0.3	0.900	0.373	4.112	4.126
EQP raw-scale SM=1.0	0.899	0.569	4.186	4.221
EQP raw-raw-scale SM=0.0	0.901	0.220	4.053	4.056
EQP raw-raw-scale SM=0.3	0.899	0.330	4.101	4.112
EQP raw-raw-scale SM=1.0	0.899	0.241	4.109	4.113

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=737), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=737) over a large number of replications (i.e., 10,000).

Table 5.11.6 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 Pain 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 raw-scale produced the smallest standard error, 0.443. That is, the difference between the mean PROMIS Pain T-score and the mean equated BPI Interference T-score based on a similar sample of 75 cases is expected to be around ±0.89 (i.e., 2 x 0.443).

Table 5.11.6: 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	-0.081	0.794	-0.071	0.555	-0.066	0.449
IRT raw-scale	-0.001	0.794	-0.007	0.551	-0.016	0.443
EQP raw-scale SM=0.0	0.384	0.812	0.386	0.563	0.390	0.454
EQP raw-scale SM=0.3	0.375	0.811	0.384	0.565	0.379	0.450
EQP raw-scale SM=1.0	0.573	0.822	0.572	0.578	0.563	0.457
EQP raw-raw-scale SM=0.0	0.221	0.793	0.221	0.552	0.220	0.445
EQP raw-raw-scale SM=0.3	0.336	0.807	0.327	0.559	0.321	0.444
EQP raw-raw-scale SM=1.0	0.243	0.814	0.235	0.565	0.234	0.447

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 31) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.12. PROMIS Anxiety and GAD-7 (Toolbox Study)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Anxiety, namely the PROMIS Anxiety (20 items) and GAD-7 (7 items). PROMIS Anxiety was scaled such that higher scores represent higher levels of Anxiety. 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.

5.12.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 100 for PROMIS Anxiety and 28 for GAD-7. Figure 5.12.1 and Figure 5.12.2 graphically display the raw summed score distributions of the two measures. Figure 5.12.3 shows the distribution for the combined. Figure 5.12.4 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 Anxiety and GAD-7 was 0.86. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anxiety and GAD-7 was 0.91. The correlations between the combined score and the measures were 0.99 and 0.91 for PROMIS Anxiety and GAD-7, respectively.

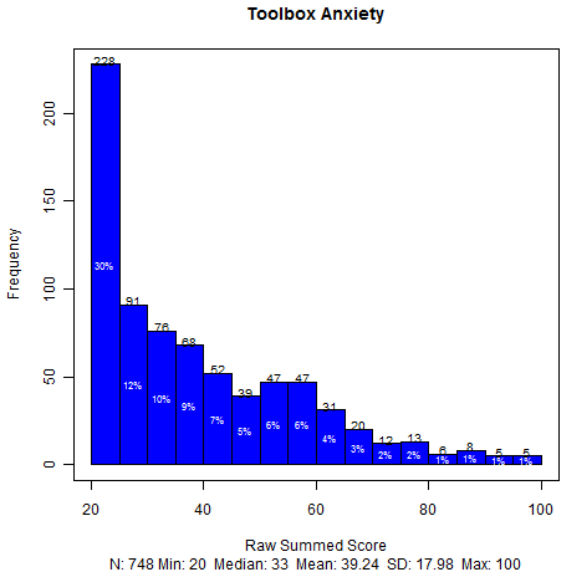

Figure 5.12.1: Raw Summed Score Distribution - PROMIS Instrument

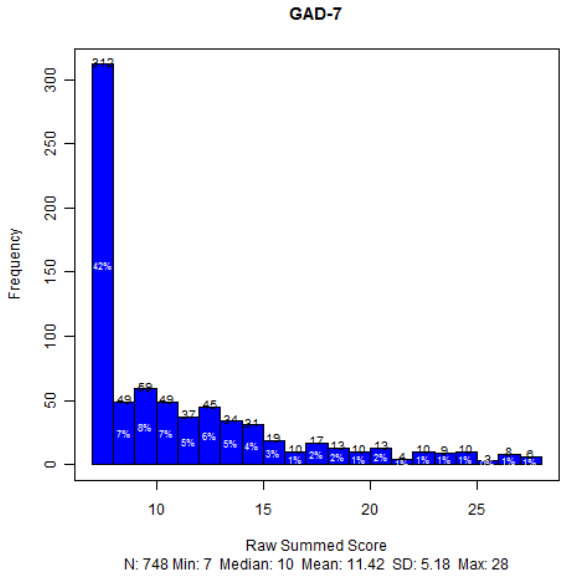

Figure 5.12.2: Raw Summed Score Distribution – Linking Instrument

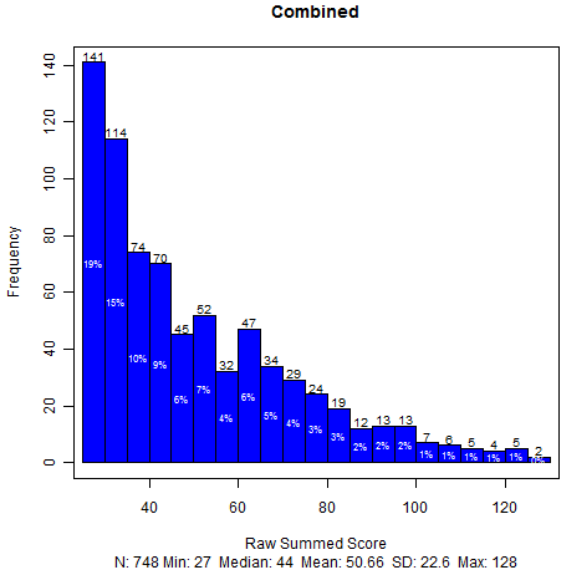

Figure 5.12.3: Raw Summed Score Distribution – Combined

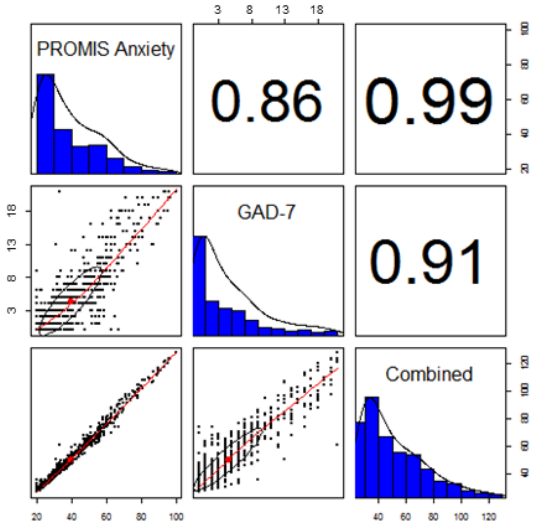

Figure 5.12.4: Scatter Plot Matrix of Raw Summed Scores

5.12.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.12.1 summarizes the results. For PROMIS Anxiety, Cronbach's alpha internal consistency reliability estimate was 0.973 and adjusted (corrected for overlap) item-total correlations ranged from 0.606 to 0.878. For GAD-7, alpha was 0.932 and adjusted item-total correlations ranged from 0.705 to 0.852. For the 27 items, alpha was 0.978 and adjusted item-total correlations ranged from 0.596 to 0.870.

Table 5.12.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Anxiety	20	0.973	0.606	0.791	0.878
GAD-7	7	0.932	0.705	0.782	0.852
Combined	27	0.978	0.596	0.781	0.870

5.12.3. 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 each of the two measures separately and on the combined. Table 5.12.2 summarizes the model fit statistics. For PROMIS Anxiety, the fit statistics were as follows: CFI = 0.983, TLI = 0.981, and RMSEA = 0.091. For GAD-7, CFI = 0.995, TLI = 0.993, and RMSEA = 0.096. For the 27 items, CFI =

0.972, TLI = 0.970, and RMSEA = 0.093. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.12.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Anxiety	20	748	0.983	0.981	0.091
GAD-7	7	748	0.995	0.993	0.096
Combined	27	748	0.972	0.970	0.093

5.12.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 27 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 27 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 20 PROMIS Anxiety items served as anchor items to transform the item parameter estimates for the GAD-7 items onto the PROMIS Anxiety 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 5.12.3 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 items were constrained to their final bank values, while the GAD-7 items were calibrated under the constraints imposed by the anchor items.

Table 5.12.3: IRT Linking Constants

	A	B
Mean/Mean	1.183	0.337
Mean/Sigma	1.283	0.257
Haebara	1.292	0.299
Stocking-Lord	1.269	0.278

The item parameter estimates for the GAD-7 items were linked to the PROMIS Anxiety metric using the transformation constants shown in Table 5.12.3. The GAD-7 item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Anxiety metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for GAD-7 as shown in Figure 5.12.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.12.6 displays the differences on the vertical axis.

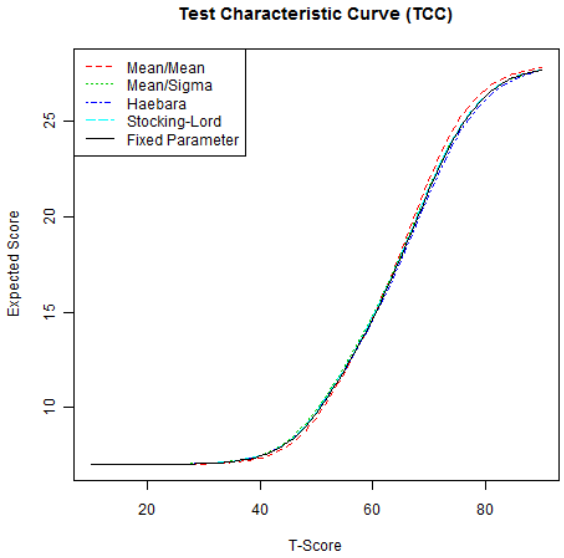

Figure 5.12.5: Test Characteristic Curves (TCC) from Different Linking Methods

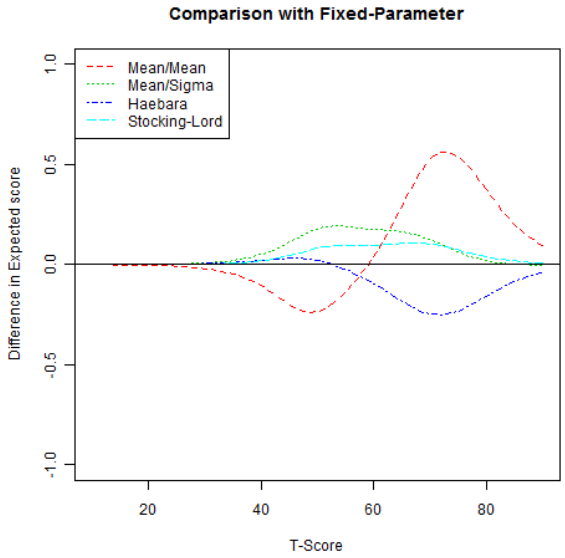

Figure 5.12.6: Difference in Test Characteristic Curves (TCC)

Table 5.12.4 shows the fixed-parameter calibration item parameter estimates for GAD-7. The marginal reliability estimate for GAD-7 based on the item parameter estimates was 0.79. The marginal reliability estimates for PROMIS Anxiety and the combined set were 0.938 and 0.945, respectively. The slope parameter estimates for GAD-7 ranged from 1.66 to 2.62 with a mean of 2.23. The slope parameter estimates for PROMIS Anxiety ranged from 1.52 to 3.88 with a mean of 2.85. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.12.7 displays the scale information functions for PROMIS Anxiety, GAD-7, and the combined set of 27. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.12.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.12.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3
2.384	0.191	1.554	2.301
2.622	0.273	1.435	2.079
2.530	0.036	1.351	1.990
2.211	0.111	1.264	1.950
1.976	0.777	1.909	2.727
1.660	0.188	1.697	2.589
2.257	0.677	1.813	2.385

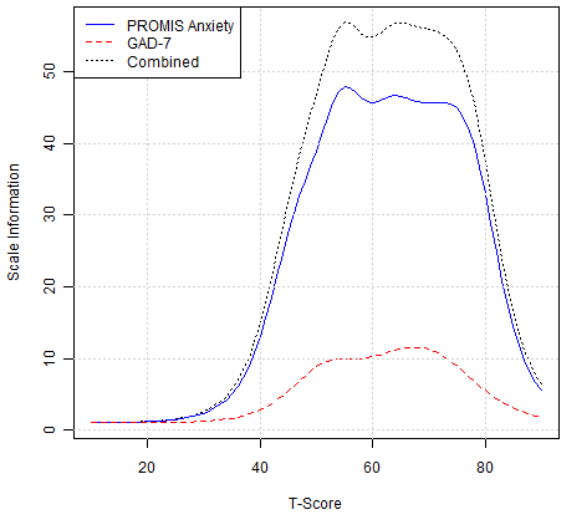

Figure 5.12.7: Comparison of Scale Information Functions

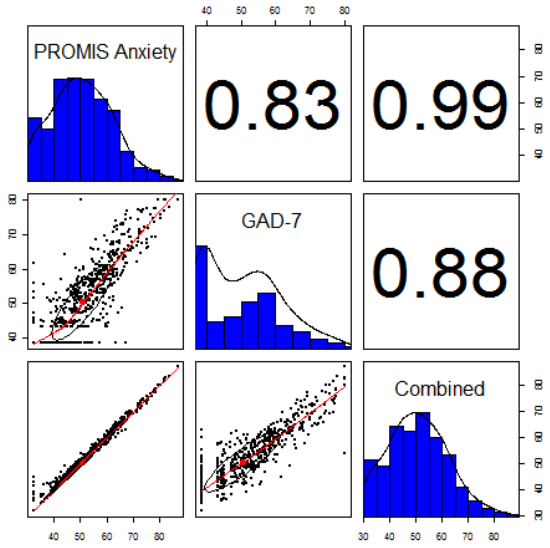

Figure 5.12.8: Comparison of IRT Scaled Scores

5.12.5. 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 GAD-7 to a scaled score on PROMIS Anxiety can be useful. Based on the GAD-7 item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 34 can be used to map simple raw summed scores from GAD-7 to T-score values linked to the PROMIS Anxiety 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.

5.12.6. Equipercentile Linking

We mapped each raw summed score point on GAD-7 to a corresponding scaled score on PROMIS Anxiety by identifying scores on PROMIS Anxiety that have the same percentile ranks as scores on GAD-7. 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 5.12.9 displays the cumulative distribution functions of the measures. Figure 5.12.10 shows the equipercentile linking functions based on raw summed scores, from GAD-7 to PROMIS Anxiety. 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 35 and Appendix Table 36 show the equipercentile crosswalk tables. The result shown in Appendix Table 35 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 36 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).

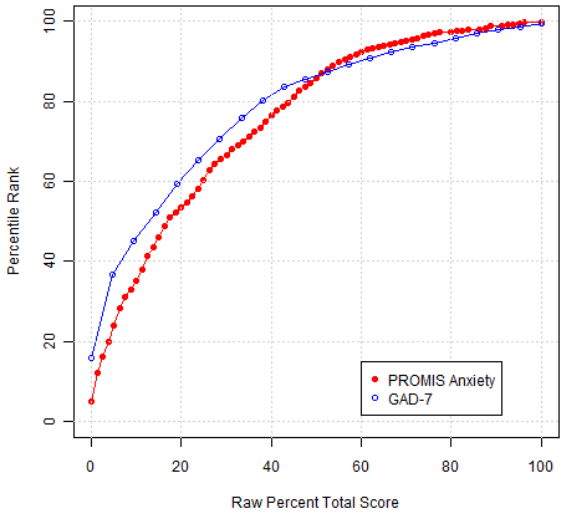

Figure 5.12.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

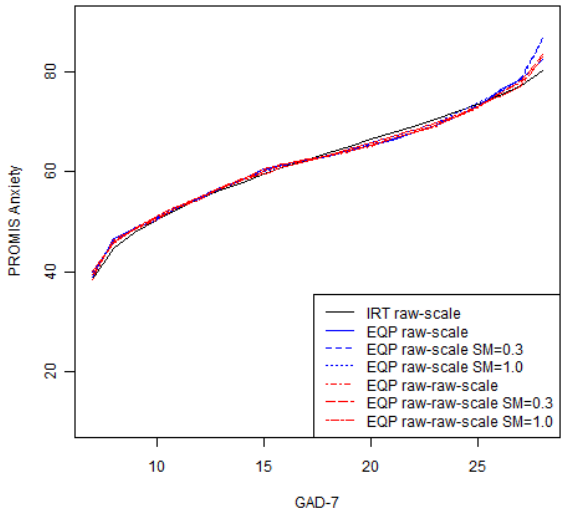

Figure 5.12.10: Equipercentile Linking Functions

5.12.7. 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 5.12.10, 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 5.12.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Anxiety T-scores and GAD-7 scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.12.10), 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 (0.829), followed by EQP raw-raw-scale SM=1.0 (0.826). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). EQP raw-raw-scale SM=1.0 yielded smallest RMSD (6.5), followed by EQP raw-rawscale SM=0.3 (6.542).

Table 5.12.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.829	0.161	6.561	6.558
IRT raw-scale	0.825	0.244	6.592	6.592
EQP raw-scale SM=0.0	0.824	-0.187	6.571	6.569
EQP raw-scale SM=0.3	0.824	-0.379	6.564	6.570
EQP raw-scale SM=1.0	0.825	-0.578	6.526	6.547
EQP raw-raw-scale SM=0.0	0.823	-0.009	6.606	6.602
EQP raw-raw-scale SM=0.3	0.825	-0.323	6.538	6.542
EQP raw-raw-scale SM=1.0	0.826	-0.545	6.481	6.500

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=748), however, subsetting out a sample was not feasible. Instead, a resampling study was used where 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 5.12.6 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 Anxiety 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, EQP raw-raw-scale SM=1.0 produced the smallest standard error, 0.71. That is, the difference between the mean PROMIS Anxiety T-score and the mean equated GAD-7 T-score based on a similar sample of 75 cases is expected to be around ±1.42 (i.e., 2 x 0.71).

Table 5.12.6: 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	0.163	1.270	0.186	0.899	0.160	0.721
IRT raw-scale	0.228	1.287	0.244	0.906	0.240	0.718
EQP raw-scale SM=0.0	-0.198	1.275	-0.206	0.900	-0.209	0.714
EQP raw-scale SM=0.3	-0.361	1.304	-0.386	0.895	-0.359	0.721
EQP raw-scale SM=1.0	-0.568	1.284	-0.586	0.885	-0.576	0.712
EQP raw-raw-scale SM=0.0	-0.009	1.281	-0.006	0.903	-0.004	0.718
EQP raw-raw-scale SM=0.3	-0.333	1.282	-0.324	0.900	-0.316	0.723
EQP raw-raw-scale SM=1.0	-0.565	1.263	-0.541	0.886	-0.537	0.710

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 34) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.13. PROMIS Anxiety and K6 (Toolbox Study)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Anxiety, namely the PROMIS Anxiety (20 items) and K6 (6 items). PROMIS Anxiety was scaled such that higher scores represent higher levels of Anxiety; for the K6, higher scores represent lower levels of Anxiety. 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.

5.13.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 100 for PROMIS Anxiety and 30 for K6. Figure 5.13.1 and Figure 5.13.2 graphically display the raw summed score distributions of the two measures. Figure 5.13.3 shows the distribution for the combined. Figure 5.13.4 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 Anxiety and K6 was -0.70. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anxiety and K6 was -0.75. The correlations between the combined score and the measures were 0.98 and 0.82 for PROMIS Anxiety and K6, respectively.

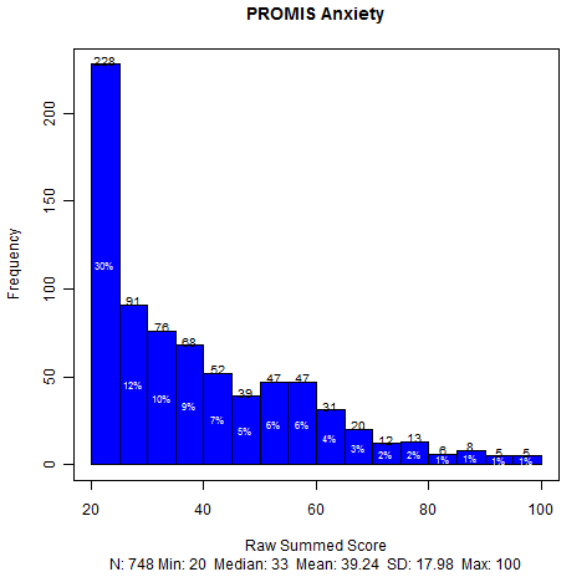

Figure 5.13.1: Raw Summed Score Distribution - PROMIS Instrument

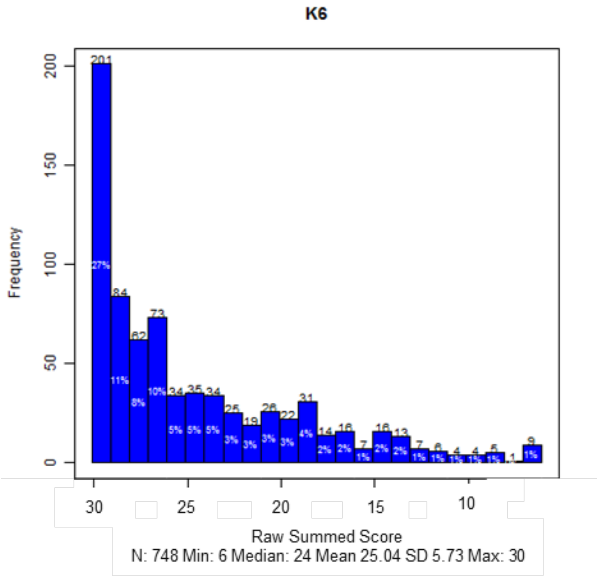

Figure 5.13.2: Raw Summed Score Distribution – Linking Instrument

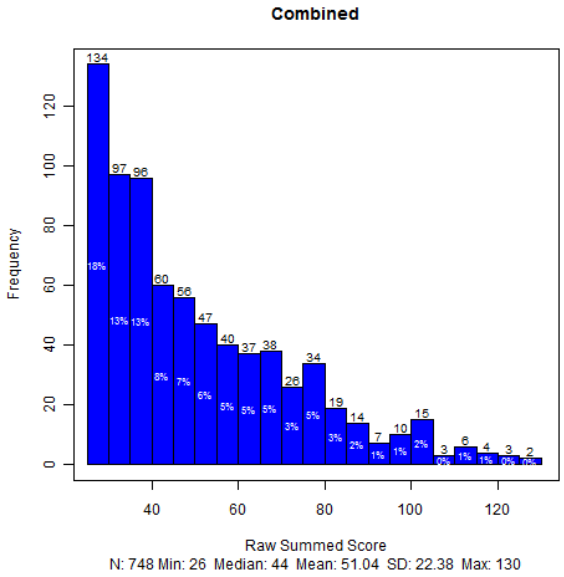

Figure 5.13.3: Raw Summed Score Distribution – Combined

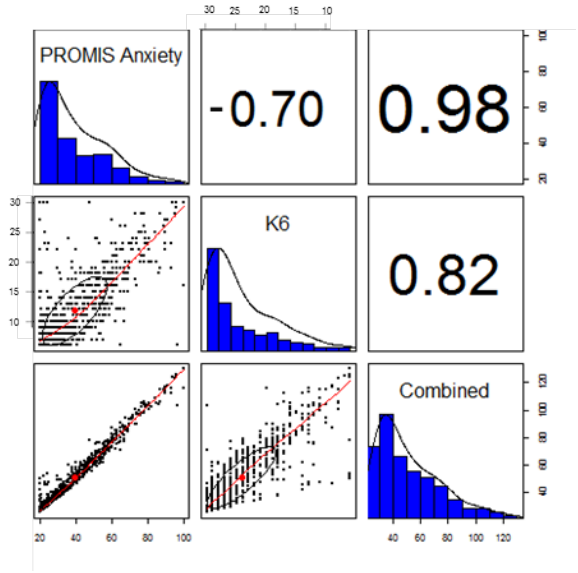

Figure 5.13.4: Scatter Plot Matrix of Raw Summed Scores

5.13.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.13.1 summarizes the results. For PROMIS Anxiety, Cronbach's alpha internal consistency reliability estimate was 0.973 and adjusted (corrected for overlap) item-total correlations ranged from 0.606 to 0.878. For K6, alpha was 0.897 and adjusted item-total correlations ranged from 0.629 to 0.812. For the 26 items, alpha was 0.972 and adjusted item-total correlations ranged from 0.542 to 0.861.

Table 5.13.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Anxiety	20	0.973	0.606	0.791	0.878
K6	6	0.897	0.629	0.723	0.812
Combined	26	0.972	0.542	0.748	0.861

5.13.3. 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 each of the two measures separately and on the combined. Table 5.13.2 summarizes the model fit statistics. For PROMIS Anxiety, the fit statistics were as follows: CFI = 0.983, TLI = 0.981, and RMSEA = 0.091. For K6, CFI = 0.979, TLI = 0.965, and RMSEA = 0.166. For the 26 items, CFI = 0.949, TLI = 0.944, and RMSEA = 0.123. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.13.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Anxiety	20	748	0.983	0.981	0.091
K6	6	748	0.979	0.965	0.166
Combined	26	748	0.949	0.944	0.123

5.13.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 26 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 26 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 20 PROMIS Anxiety items served as anchor items to transform the item parameter estimates for the K6 items onto the PROMIS Anxiety 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 5.13.3 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 items were constrained to their final bank values, while theK6 items were calibrated under the constraints imposed by the anchor items.

Table 5.13.3: IRT Linking Constants

	A	B
Mean/Mean	1.172	0.299
Mean/Sigma	1.267	0.219
Haebara	1.274	0.262
Stocking-Lord	1.253	0.240

The item parameter estimates for the K6 items were linked to the PROMIS Anxiety metric using the transformation constants shown in Table 5.13.3. The K6 item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Anxiety metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for K6 as shown in Figure 5.13.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.13.6 displays the differences on the vertical axis.

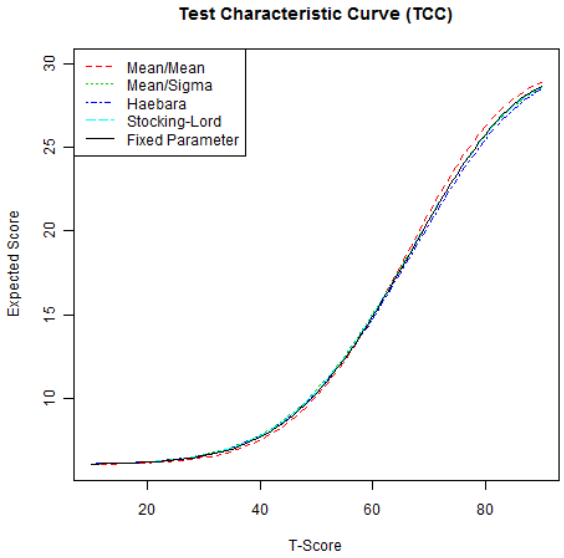

Figure 5.13.5: Test Characteristic Curves (TCC) from Different Linking Methods

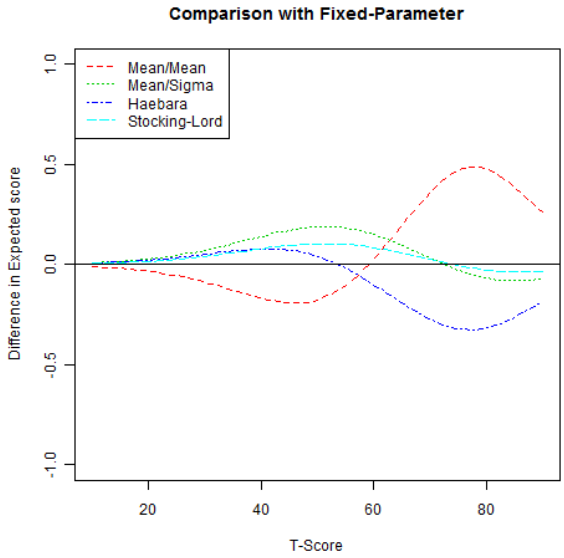

Figure 5.13.6: Difference in Test Characteristic Curves (TCC)

Table 5.13.4 shows the fixed-parameter calibration item parameter estimates for K6. The marginal reliability estimate for K6 based on the item parameter estimates was 0.729. The marginal reliability estimates for PROMIS Anxiety and the combined set were 0.938 and 0.947, respectively. The slope parameter estimates for K6 ranged from 1.24 to 1.93 with a mean of 1.51. The slope parameter estimates for PROMIS Anxiety ranged from 1.52 to 3.88 with a mean of 2.85. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.13.7 displays the scale information functions for PROMIS Anxiety, K6, and the combined set of 26. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.13.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.13.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
1.236	-0.699	0.843	1.939	2.592
1.427	0.256	1.233	2.111	2.696
1.384	-0.319	0.970	2.073	3.019
1.707	0.607	1.437	2.137	2.803
1.387	-0.442	0.877	1.972	3.027
1.926	0.573	1.306	2.030	2.810

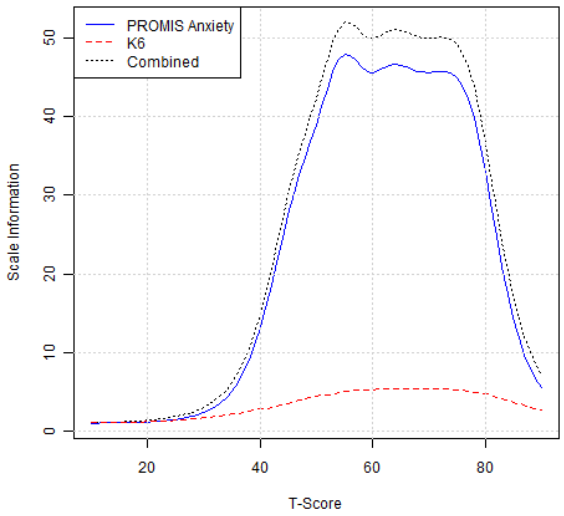

Figure 5.13.7: Comparison of Scale Information Functions

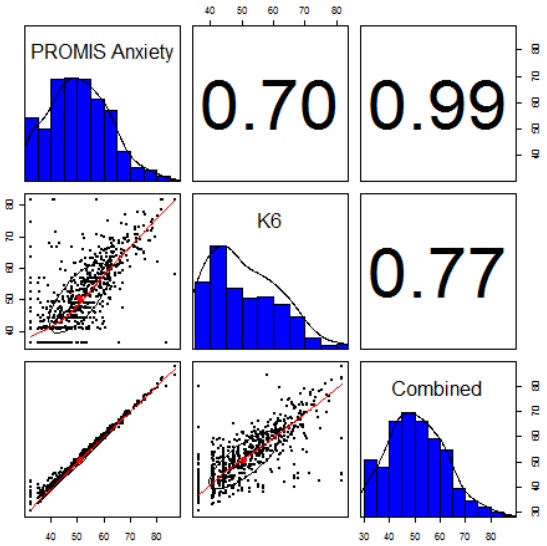

Figure 5.13.8: Comparison of IRT Scaled Scores

5.13.5. 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 K6 to a scaled score on PROMIS Anxiety can be useful. Based on the K6 item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 37 can be used to map simple raw summed scores from K6 to T-score values linked to the PROMIS Anxiety 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.

5.13.6. Equipercentile Linking

We mapped each raw summed score point on K6 to a corresponding scaled score on PROMIS Anxiety by identifying scores on PROMIS Anxiety that have the same percentile ranks as scores on K6. 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 5.13.9 displays the cumulative distribution functions of the measures. Figure 5.13.10 shows the equipercentile linking functions based on raw summed scores, from K6 to PROMIS Anxiety. 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 38 and Appendix Table 39 show the equipercentile crosswalk tables. The result shown in Appendix Table 38 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 39 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).

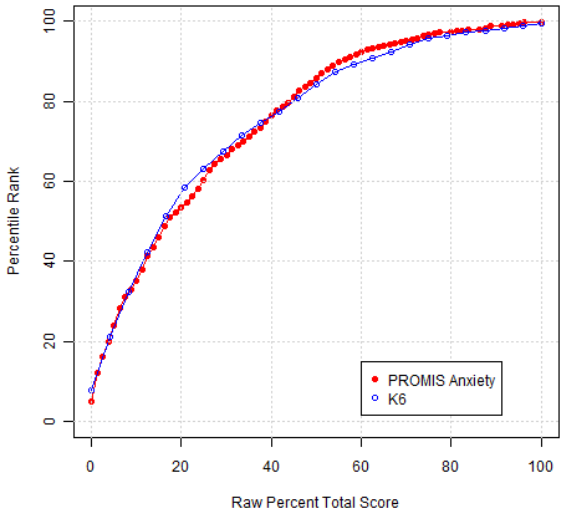

Figure 5.13.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

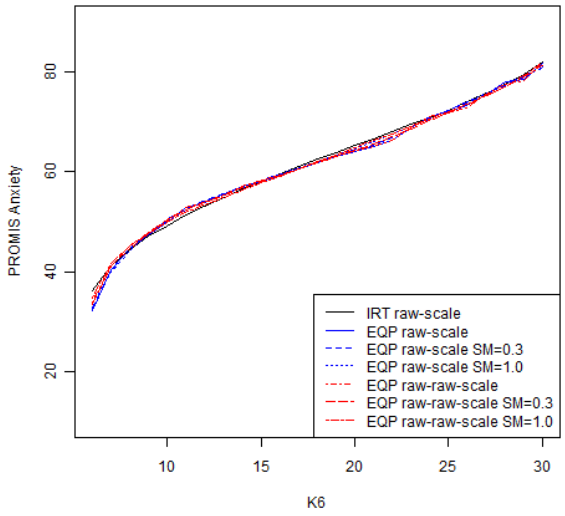

Figure 5.13.10: Equipercentile Linking Functions

5.13.7. 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 5.13.10, 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 5.13.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Anxiety T-

scores and K6 scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.13.10), 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 (0.696), followed by EQP raw-scale SM=1.0 (0.674). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (8.759), followed by EQP raw-raw-scale SM=1.0 (9.067). The low correlations indicate the two measures may be significantly different from each other. The disattenuated correlation of 0.75 was still very low (less than 0.80). Caution should be demonstrated when using these linking tables.

Table 5.13.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.696	-0.093	8.764	8.759
IRT raw-scale	0.669	-0.001	9.076	9.070
EQP raw-scale SM=0.0	0.672	0.527	9.330	9.339
EQP raw-scale SM=0.3	0.673	0.441	9.264	9.268
EQP raw-scale SM=1.0	0.674	0.351	9.235	9.235
EQP raw-raw-scale SM=0.0	0.670	0.210	9.176	9.172
EQP raw-raw-scale SM=0.3	0.671	0.221	9.206	9.203
EQP raw-raw-scale SM=1.0	0.671	0.037	9.073	9.067

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=748), however, subsetting out a sample was not feasible. Instead, a resampling study was used where 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 5.13.6 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 Anxiety 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, 0.965. That is, the difference between the mean PROMIS Anxiety T-score and the mean equated K6 T-score based on a similar sample of 75 cases is expected to be around ±1.93 (i.e., 2 x 0.965).

Table 5.13.6: 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	-0.103	1.733	-0.112	1.194	-0.078	0.965
IRT raw-scale	-0.002	1.765	0.003	1.252	-0.018	1.007
EQP raw-scale SM=0.0	0.514	1.821	0.532	1.261	0.527	1.004
EQP raw-scale SM=0.3	0.439	1.831	0.446	1.268	0.443	1.011
EQP raw-scale SM=1.0	0.356	1.800	0.351	1.276	0.332	1.007
EQP raw-raw-scale SM=0.0	0.191	1.801	0.236	1.263	0.218	1.001
EQP raw-raw-scale SM=0.3	0.188	1.812	0.236	1.261	0.234	1.011
EQP raw-raw-scale SM=1.0	0.032	1.777	0.041	1.234	0.035	1.002

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 37) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.14. PROMIS Anxiety and MASQ (Toolbox Study)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Anxiety, namely the PROMIS Anxiety (20 items) and MASQ (28 items). PROMIS Anxiety was scaled such that higher scores represent higher levels of Anxiety. 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.

5.14.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 100 for PROMIS Anxiety and 140 for MASQ. Figure 5.14.1 and Figure 5.14.2 graphically display the raw summed score distributions of the two measures. Figure 5.14.3 shows the distribution for the combined. Figure 5.14.4 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 Anxiety and MASQ was 0.82. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anxiety and MASQ was 0.85. The correlations between the combined score and the measures were 0.96 and 0.95 for PROMIS Anxiety and MASQ, respectively.

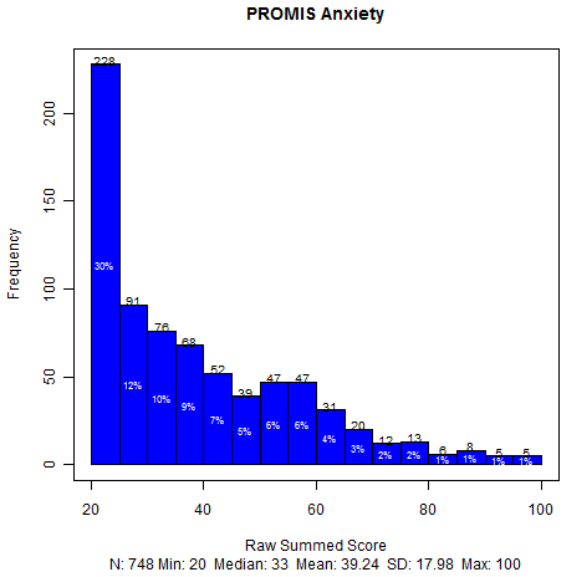

Figure 5.14.1: Raw Summed Score Distribution - PROMIS Instrument

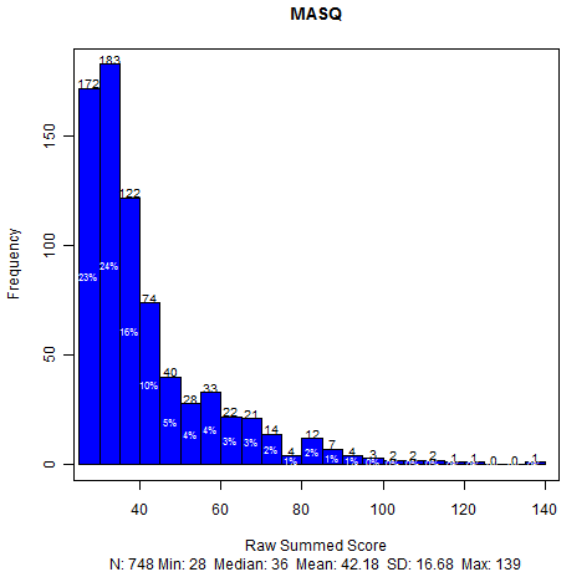

Figure 5.14.2: Raw Summed Score Distribution – Linking Instrument

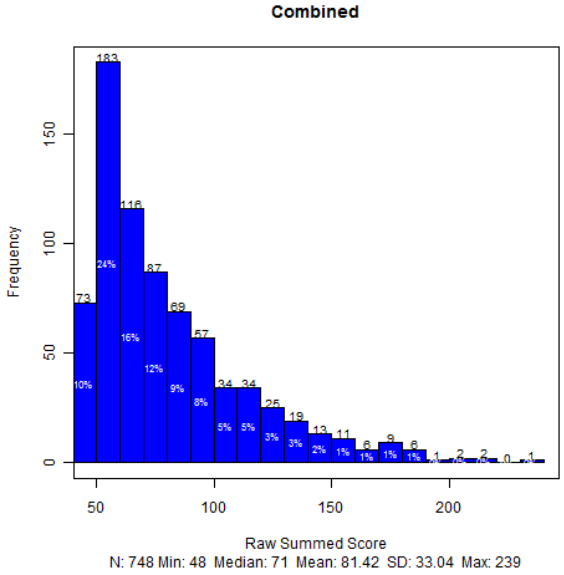

Figure 5.14.3: Raw Summed Score Distribution – Combined

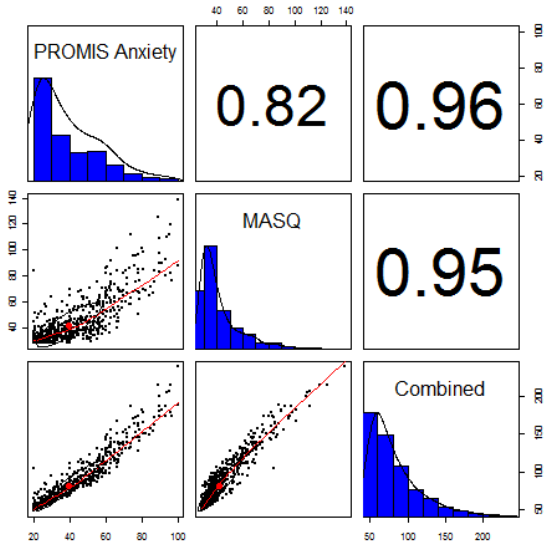

Figure 5.14.4: Scatter Plot Matrix of Raw Summed Scores

5.14.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.14.1 summarizes the results. For PROMIS Anxiety, Cronbach's alpha internal consistency reliability estimate was 0.973 and adjusted (corrected for overlap) item-total correlations ranged from 0.606 to 0.878. For MASQ, alpha was 0.958 and adjusted item-total correlations ranged from 0.500 to 0.818. For the 48 items, alpha was 0.979 and adjusted item-total correlations ranged from 0.465 to 0.859.

Table 5.14.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Anxiety	20	0.973	0.606	0.791	0.878
MASQ	28	0.958	0.500	0.665	0.818
Combined	48	0.979	0.465	0.691	0.859

5.14.3. 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 each of the two measures separately and on the combined. Table 5.14.2 summarizes the model fit statistics. For PROMIS Anxiety, the fit statistics were as follows: CFI = 0.983, TLI = 0.981, and RMSEA = 0.091. For MASQ, CFI = 0.935, TLI = 0.930, and RMSEA = 0.08. For the 48 items, CFI = 0.939,

TLI = 0.936, and RMSEA = 0.078. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.14.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Anxiety	20	748	0.983	0.981	0.091
MASQ	28	748	0.935	0.930	0.080
Combined	48	748	0.939	0.936	0.078

5.14.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 48 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 48 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 20 PROMIS Anxiety items served as anchor items to transform the item parameter estimates for the MASQ items onto the PROMIS Anxiety 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 5.14.3 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 items were constrained to their final bank values, while the MASQ items were calibrated under the constraints imposed by the anchor items.

Table 5.14.3: IRT Linking Constants

	A	B
Mean/Mean	1.224	0.658
Mean/Sigma	1.355	0.591
Haebara	1.363	0.634
Stocking-Lord	1.334	0.612

The item parameter estimates for the MASQ items were linked to the PROMIS Anxiety metric using the transformation constants shown in Table 5.14.3. The MASQ item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Anxiety metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for MASQ as shown in Figure 5.14.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.14.6 displays the differences on the vertical axis.

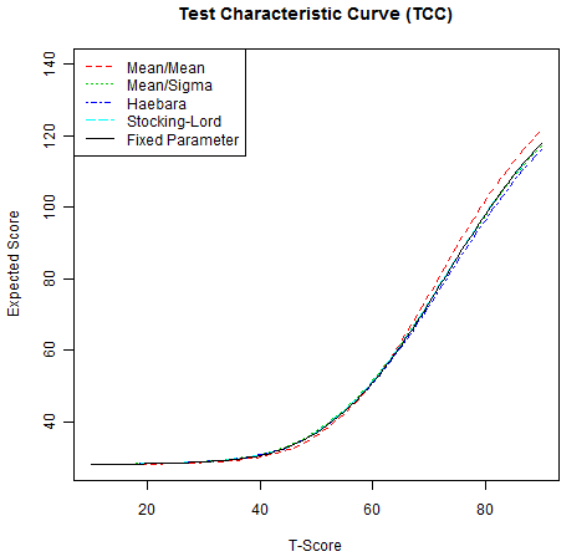

Figure 5.14.5: Test Characteristic Curves (TCC) from Different Linking Methods

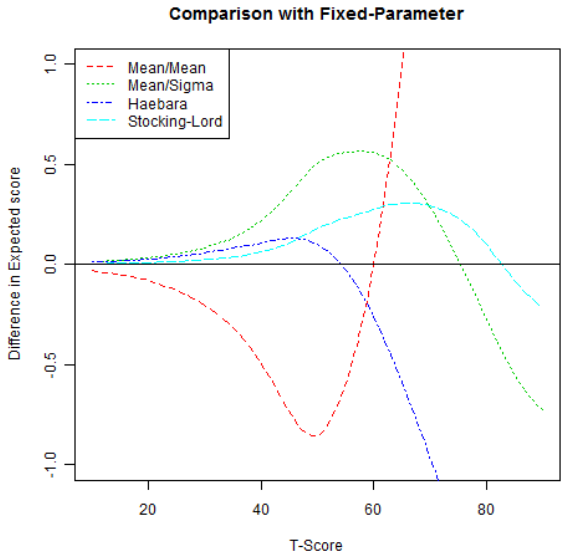

Figure 5.14.6: Difference in Test Characteristic Curves (TCC)

Table 5.14.4 shows the fixed-parameter calibration item parameter estimates for MASQ. The marginal reliability estimate for MASQ based on the item parameter estimates was 0.881. The marginal reliability estimates for PROMIS Anxiety and the combined set were 0.938 and 0.954, respectively. The slope parameter estimates for MASQ ranged from 0.878 to 3.12 with a mean of 1.63. The slope parameter estimates for PROMIS Anxiety ranged from 1.52 to 3.88 with a mean of 2.85. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.14.7 displays the scale information functions for PROMIS Anxiety, MASQ, and the combined set of 48. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.14.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.14.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
0.969	0.283	1.768	3.280	4.388
2.112	0.475	1.665	2.323	3.170
1.590	0.945	2.082	2.833	3.774
0.878	1.280	3.054	4.452	5.681
2.848	0.011	1.140	1.871	2.651
1.131	0.723	2.121	3.217	4.696
3.123	0.030	1.137	1.805	2.636
1.627	1.450	2.383	3.303	4.423
1.464	1.356	2.523	3.263	4.257
1.115	0.892	2.111	3.098	4.273
1.286	0.383	1.910	2.973	4.334
1.245	1.407	2.564	3.606	4.827
2.243	0.413	1.525	2.174	2.997
2.042	1.320	2.187	2.873	3.674
1.601	1.733	2.646	3.719	4.320
2.423	0.058	1.205	1.895	2.610
1.446	0.903	2.163	2.846	4.390
1.293	0.836	2.228	3.436	4.464
1.488	1.264	2.568	3.576	4.251
1.656	2.089	2.821	3.838	4.845
1.265	1.089	2.341	3.504	4.910
1.090	0.705	2.044	3.167	4.232
1.765	1.775	2.567	3.342	3.853
1.796	1.053	2.063	2.769	3.699
1.269	1.053	2.227	3.341	4.241
2.392	0.486	1.393	2.166	2.939
1.311	0.106	1.541	2.555	3.711
1.043	0.295	1.711	2.558	4.063

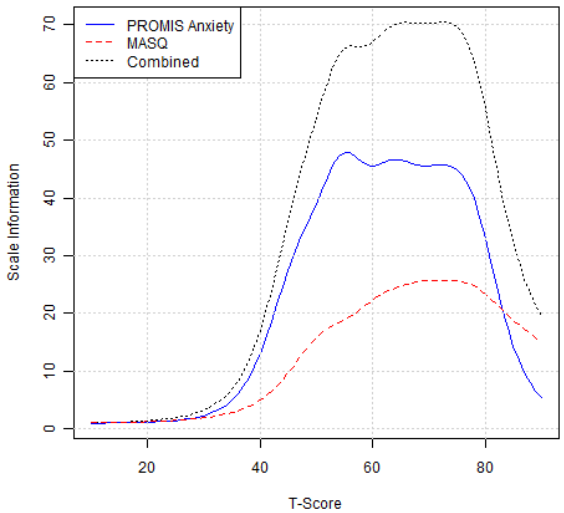

Figure 5.14.7: Comparison of Scale Information Functions

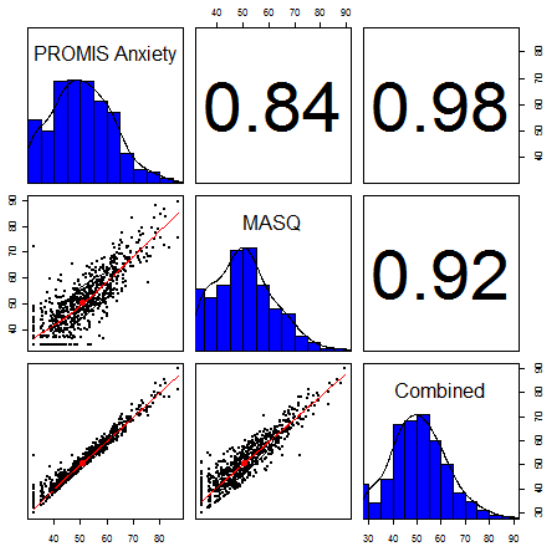

Figure 5.14.8: Comparison of IRT Scaled Scores

5.14.5. 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 MASQ to a scaled score on PROMIS Anxiety can be useful. Based on the MASQ item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 40 can be used to map simple raw summed scores from MASQ to T-score values linked to the PROMIS Anxiety 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.

5.14.6. Equipercentile Linking

We mapped each raw summed score point on MASQ to a corresponding scaled score on PROMIS Anxiety by identifying scores on PROMIS Anxiety that have the same percentile ranks as scores on MASQ. 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 5.14.9 displays the cumulative distribution functions of the measures. Figure 5.14.10 shows the equipercentile linking functions based on raw summed scores, from MASQ to PROMIS Anxiety. 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 41 and Appendix Table 42 show the equipercentile crosswalk tables. The result shown in Appendix Table 41 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 42 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).

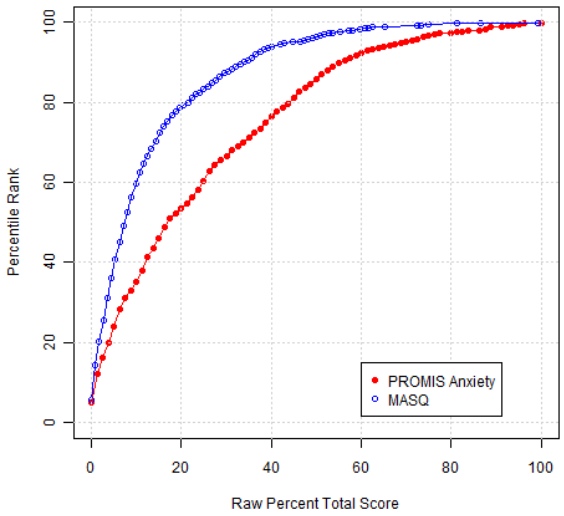

Figure 5.14.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

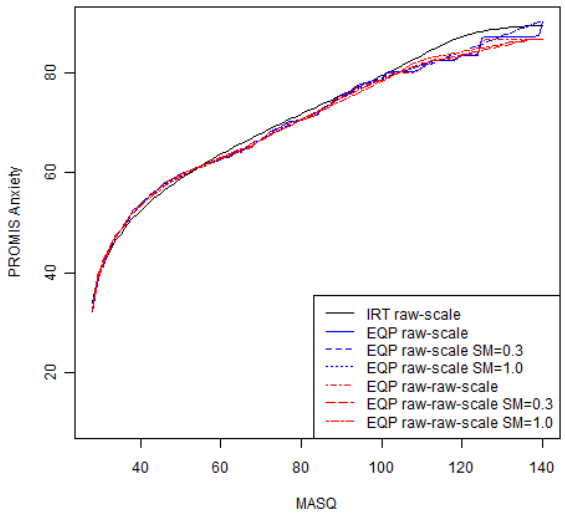

Figure 5.14.10: Equipercentile Linking Functions

5.14.7. 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 5.14.10, 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 5.14.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Anxiety T-scores and MASQ scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.14.10), 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 (0.838), followed by IRT raw-scale (0.802). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (6.46), followed by IRT raw-scale (7.169).

Table 5.14.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.838	0.101	6.463	6.460
IRT raw-scale	0.802	0.440	7.160	7.169
EQP raw-scale SM=0.0	0.797	0.340	7.302	7.305
EQP raw-scale SM=0.3	0.798	0.283	7.256	7.256
EQP raw-scale SM=1.0	0.798	0.387	7.280	7.286
EQP raw-raw-scale SM=0.0	0.798	0.198	7.204	7.202
EQP raw-raw-scale SM=0.3	0.798	0.290	7.240	7.241
EQP raw-raw-scale SM=1.0	0.798	0.239	7.189	7.188

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=748), however, subsetting out a sample was not feasible. Instead, a resampling study was used where 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 5.14.6 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 Anxiety 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, 0.7. That is, the difference between the mean PROMIS Anxiety T-score and the mean equated MASQ T-score based on a similar sample of 75 cases is expected to be around ±1.4 (i.e., 2 x 0.70).

Table 5.14.6: 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	0.135	1.278	0.097	0.878	0.103	0.700
IRT raw-scale	0.422	1.409	0.435	0.988	0.431	0.779
EQP raw-scale SM=0.0	0.339	1.431	0.335	0.997	0.337	0.795
EQP raw-scale SM=0.3	0.267	1.423	0.268	0.993	0.278	0.795
EQP raw-scale SM=1.0	0.380	1.431	0.384	0.992	0.383	0.796
EQP raw-raw-scale SM=0.0	0.185	1.394	0.200	0.981	0.196	0.775
EQP raw-raw-scale SM=0.3	0.294	1.431	0.274	0.985	0.293	0.782
EQP raw-raw-scale SM=1.0	0.241	1.409	0.227	0.981	0.253	0.789

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 40) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.15. PROMIS Depression and CES-D (Toolbox Study)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression (20 items) and CES-D (20 items). PROMIS Depression was scaled such that higher scores represent higher levels of Depression. 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.

5.15.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 100 for PROMIS Depression and 60 for CES-D. Figure 5.15.1 and Figure 5.15.2 graphically display the raw summed score distributions of the two measures. Figure 5.15.3 shows the distribution for the combined. Figure 5.15.4 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 CES-D was 0.88. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and CES-D was 0.92. The correlations between the combined score and the measures were 0.98 and 0.95 for PROMIS Depression and CES-D, respectively.

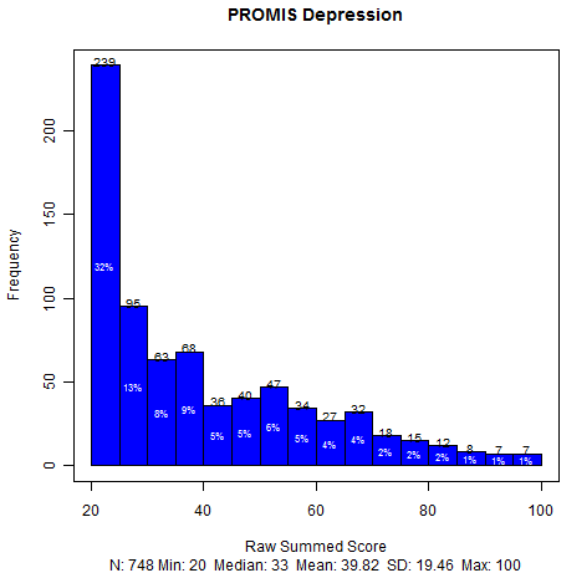

Figure 5.15.1: Raw Summed Score Distribution - PROMIS Instrument

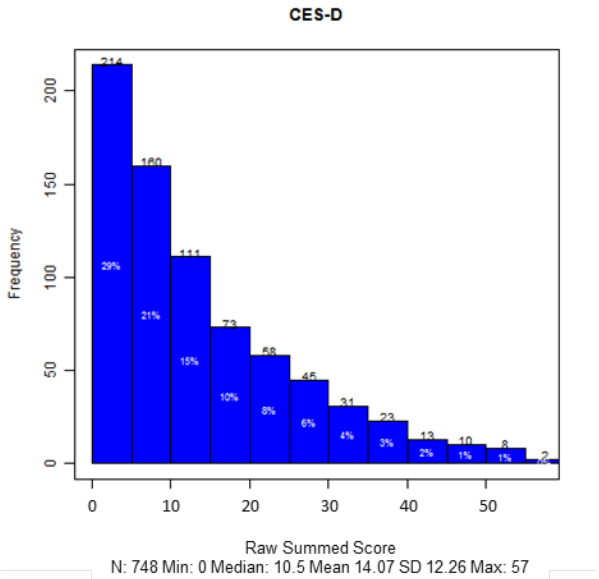

Figure 5.15.2: Raw Summed Score Distribution – Linking Instrument

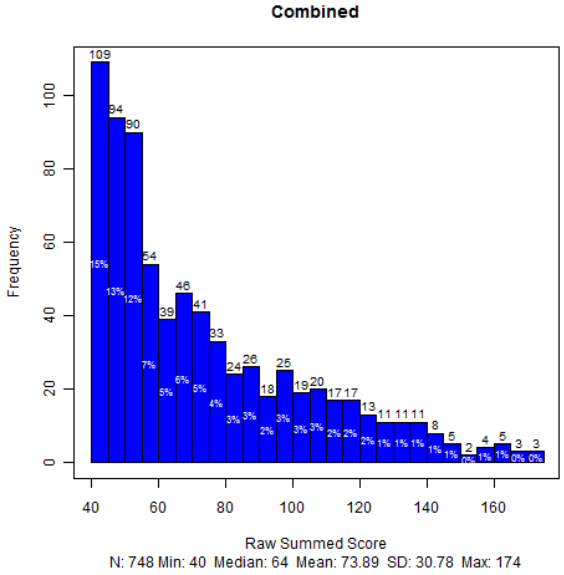

Figure 5.15.3: Raw Summed Score Distribution – Combined

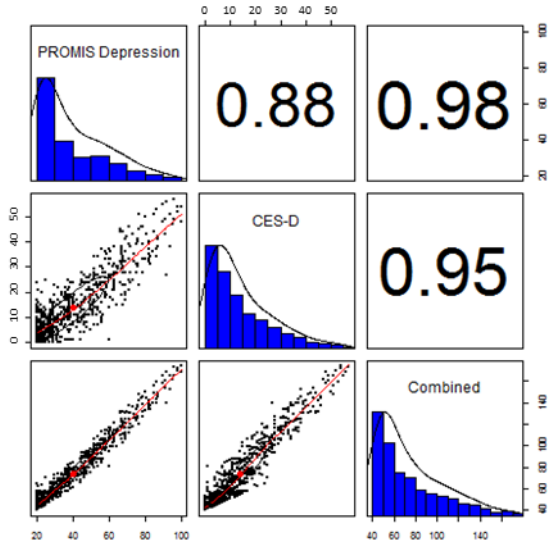

Figure 5.15.4: Scatter Plot Matrix of Raw Summed Scores

5.15.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.15.1 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 CES-D, alpha was 0.935 and adjusted item-total correlations ranged from 0.429 to 0.819. For the 40 items, alpha was 0.979 and adjusted item-total correlations ranged from 0.410 to 0.872.

Table 5.15.1: 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
CES-D	20	0.935	0.429	0.634	0.819
Combined	40	0.979	0.410	0.722	0.872

5.15.3. 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 each of the two measures separately and on the combined. Table 5.15.2 summarizes the model fit statistics. For PROMIS Depression, the fit statistics were as follows: CFI = 0.988, TLI = 0.986, and RMSEA = 0.089. For CES-D, CFI = 0.915, TLI = 0.906, and RMSEA = 0.120. For the 40 items,

CFI = 0.954, TLI = 0.951, and RMSEA = 0.093. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.15.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Depression	20	748	0.988	0.986	0.089
CES-D	20	748	0.915	0.906	0.120
Combined	40	748	0.954	0.951	0.093

5.15.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 40 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 40 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 20 PROMIS Depression items served as anchor items to transform the item parameter estimates for the CES-D 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 5.15.3 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 items were constrained to their final bank values, while the CES-D items were calibrated under the constraints imposed by the anchor items.

Table 5.15.3: IRT Linking Constants

	A	B
Mean/Mean	1.126	0.469
Mean/Sigma	1.205	0.419
Haebara	1.206	0.444
Stocking-Lord	1.191	0.433

The item parameter estimates for the CES-D items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.15.3. The CES-D 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 CES-D as shown in Figure 5.15.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.15.6 displays the differences on the vertical axis.

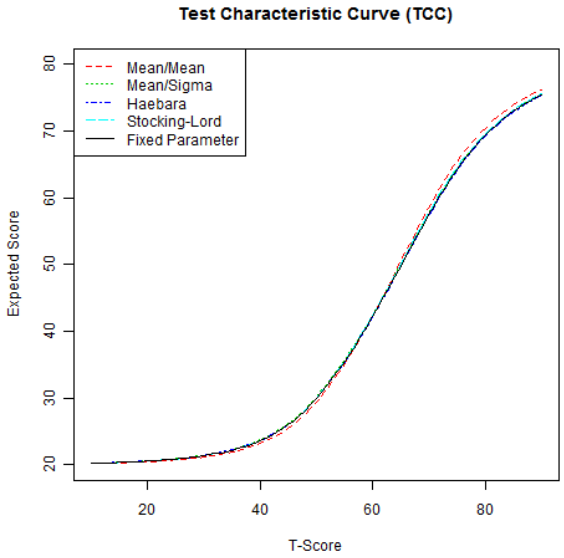

Figure 5.15.5: Test Characteristic Curves (TCC) from Different Linking Methods

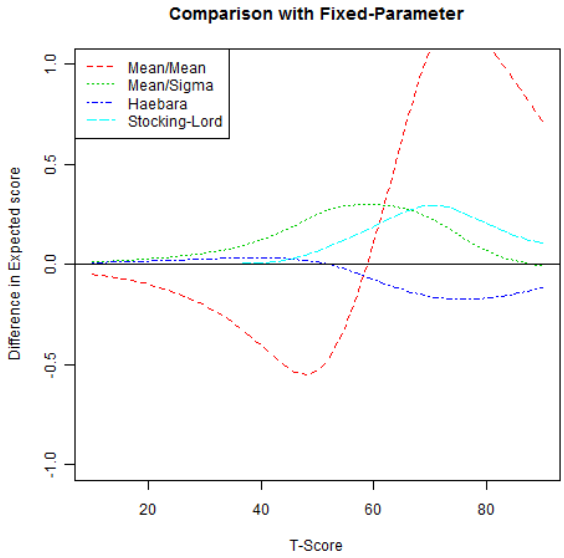

Figure 5.15.6: Difference in Test Characteristic Curves (TCC)

Table 5.15.4 shows the fixed-parameter calibration item parameter estimates for CES-D. The marginal reliability estimate for CES-D based on the item parameter estimates was 0.881. The marginal reliability estimates for PROMIS Depression and the combined set were 0.929 and 0.952, respectively. The slope parameter estimates for CES-D ranged from 0.794 to 3.03 with a mean of 1.66. 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 5.15.7 displays the scale information functions for PROMIS Depression, CES-D, and the combined set of 40. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.15.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.15.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3
1.368	0.692	2.164	3.493
0.927	1.169	2.622	4.367
2.265	0.764	1.571	2.411
0.803	0.083	1.477	2.736
1.410	0.295	1.698	2.921
2.992	0.332	1.215	2.013
1.463	0.075	1.341	2.368
0.794	-0.842	0.703	2.137
2.560	0.627	1.519	2.166
1.920	0.871	1.877	2.752
1.176	-0.335	1.011	2.162
1.697	-0.137	1.007	1.968
1.046	0.090	1.932	3.499
1.993	0.185	1.226	2.077
1.099	0.926	2.725	4.056
1.841	0.050	0.987	1.983
1.471	1.234	2.154	3.296
3.026	0.064	1.239	1.989
1.697	0.940	2.132	2.955
1.682	0.015	1.324	2.530

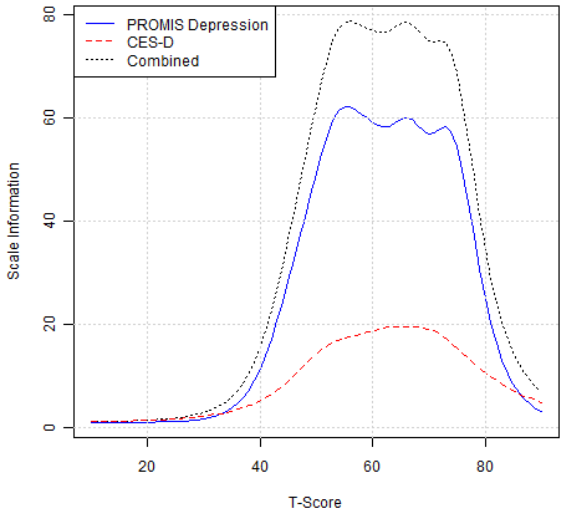

Figure 5.15.7: Comparison of Scale Information Functions

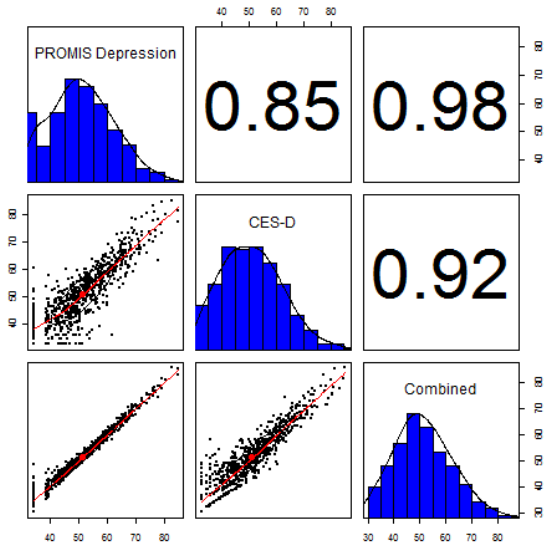

Figure 5.15.8: Comparison of IRT Scaled Scores

5.15.5. 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 CES-D to a scaled score on PROMIS

Depression can be useful. Based on the CES-D item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 43 can be used to map simple raw summed scores from CES-D 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.

5.15.6. Equipercentile Linking

We mapped each raw summed score point on CES-D to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on CES-D. 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 5.15.9 displays the cumulative distribution functions of the measures. Figure 5.15.10 shows the equipercentile linking functions based on raw summed scores, from CES-D 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 44 and Appendix Table 45 show the equipercentile crosswalk tables. The result shown in Appendix Table 44 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 45 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).

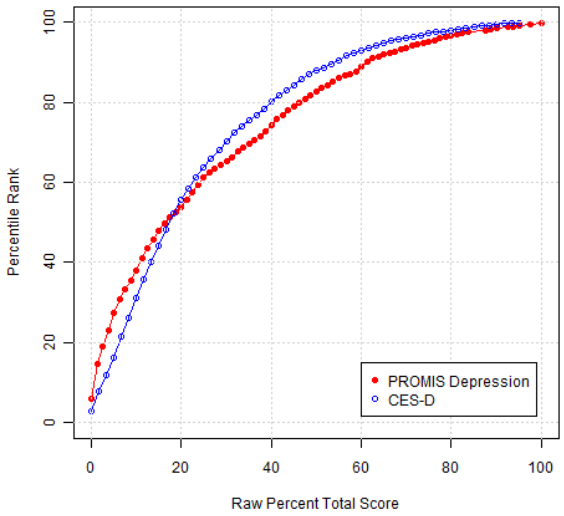

Figure 5.15.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

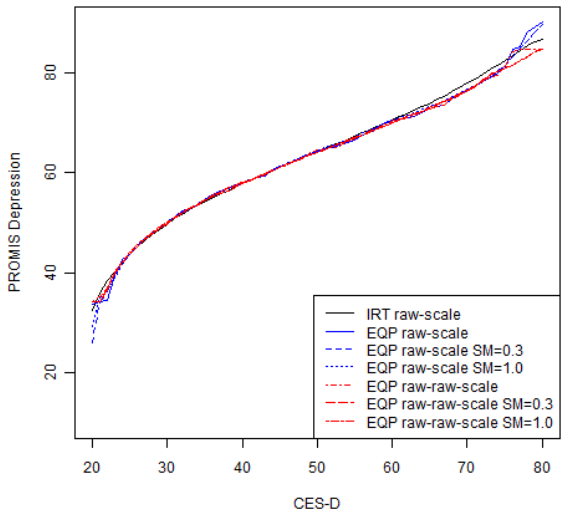

Figure 5.15.10: Equipercentile Linking Functions

5.15.7. 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 5.15.10, 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 5.15.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Depression T-scores and CES-D scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.15.10), 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 (0.853), followed by IRT raw-scale (0.814). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.987), followed by IRT raw-scale (6.657).

Table 5.15.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.853	0.207	5.988	5.987
IRT raw-scale	0.814	0.107	6.661	6.657
EQP raw-scale SM=0.0	0.804	0.142	6.847	6.843
EQP raw-scale SM=0.3	0.793	0.554	7.317	7.333
EQP raw-scale SM=1.0	0.804	0.397	6.994	7.000
EQP raw-raw-scale SM=0.0	0.810	0.043	6.686	6.681
EQP raw-raw-scale SM=0.3	0.808	0.103	6.733	6.729
EQP raw-raw-scale SM=1.0	0.806	0.060	6.749	6.745

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=748), however, subsetting out a sample was not feasible. Instead, a resampling study was used where 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 5.15.6 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, 0.658. That is, the difference between the mean PROMIS Depression T-score and the mean equated CES-D T-score based on a similar sample of 75 cases is expected to be around ±1.32 (i.e., 2 x 0.658).

Table 5.15.6: 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	0.214	1.169	0.203	0.814	0.196	0.658
IRT raw-scale	0.086	1.302	0.103	0.914	0.109	0.730
EQP raw-scale SM=0.0	0.151	1.353	0.150	0.933	0.141	0.754
EQP raw-scale SM=0.3	0.556	1.451	0.548	0.999	0.533	0.808
EQP raw-scale SM=1.0	0.419	1.360	0.381	0.945	0.396	0.767
EQP raw-raw-scale SM=0.0	0.023	1.318	0.052	0.894	0.038	0.736
EQP raw-raw-scale SM=0.3	0.105	1.319	0.118	0.920	0.114	0.743
EQP raw-raw-scale SM=1.0	0.042	1.322	0.051	0.918	0.062	0.736

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 43) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.16. PROMIS Depression and PHQ-9 (Toolbox Study)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression (20 items) and PHQ-9 (9 items). PROMIS Depression was scaled such that higher scores represent higher levels of Depression. 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.

5.16.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 100 for PROMIS Depression and 27 for PHQ-9. Figure 5.16.1 and Figure 5.16.2 graphically display the raw summed score distributions of the two measures. Figure 5.16.3 shows the distribution for the combined. Figure 5.16.4 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-9 was 0.84. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and PHQ-9 was 0.89. The correlations between the combined score and the measures were 0.99 and 0.90 for PROMIS Depression and PHQ-9, respectively.

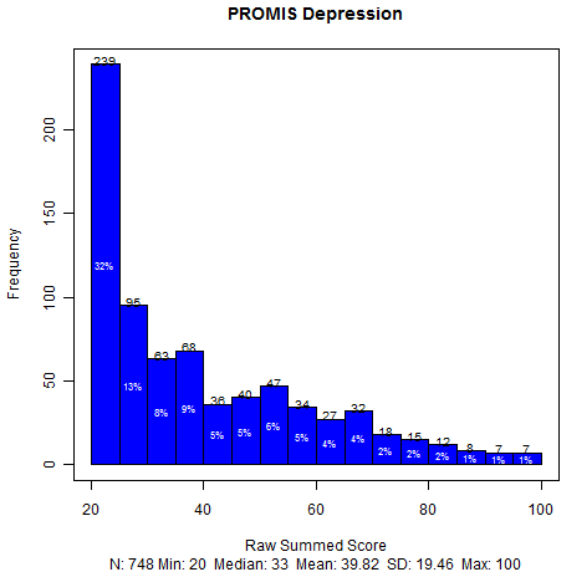

Figure 5.16.1: Raw Summed Score Distribution - PROMIS Instrument

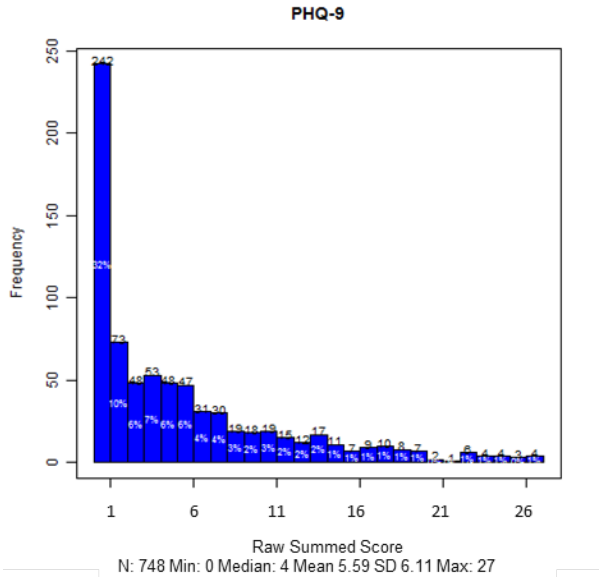

Figure 5.16.2: Raw Summed Score Distribution – Linking Instrument

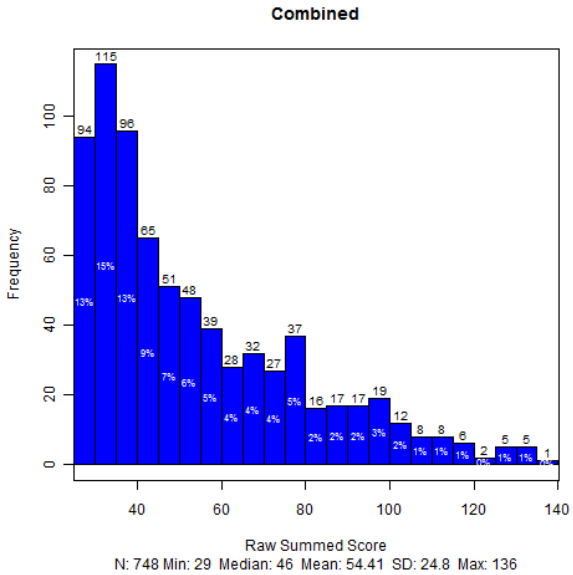

Figure 5.16.3: Raw Summed Score Distribution – Combined

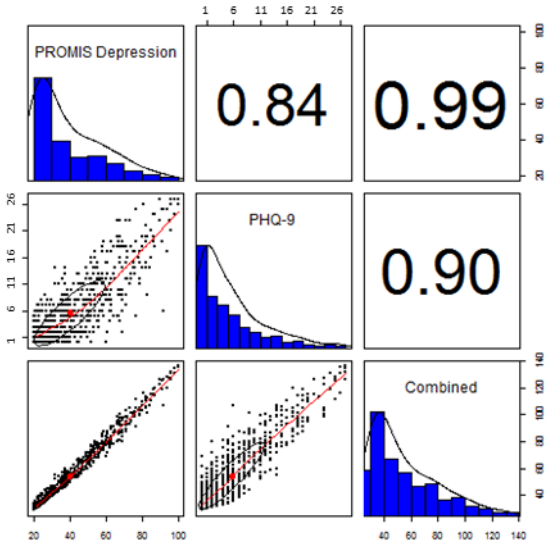

Figure 5.16.4: Scatter Plot Matrix of Raw Summed Scores

5.16.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.16.1 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-9, alpha was 0.912 and adjusted item-total correlations ranged from 0.581 to 0.796. For the 29 items, alpha was 0.979 and adjusted item-total correlations ranged from 0.601 to 0.875.

Table 5.16.1: 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-9	9	0.912	0.581	0.705	0.796
Combined	29	0.979	0.601	0.778	0.875

5.16.3. 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 each of the two measures separately and on the combined. Table 5.16.2 summarizes the model fit statistics. For PROMIS Depression, the fit statistics were as follows: CFI = 0.988, TLI = 0.986, and RMSEA = 0.089. For PHQ-9, CFI = 0.985, TLI = 0.98, and RMSEA = 0.091. For the 29 items,

CFI = 0.977, TLI = 0.975, and RMSEA = 0.087. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.16.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Depression	20	748	0.988	0.986	0.089
PHQ-9	9	748	0.985	0.980	0.091
Combined	29	748	0.977	0.975	0.087

5.16.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 29 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 29 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 20 PROMIS Depression items served as anchor items to transform the item parameter estimates for the PHQ-9 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 5.16.3 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 items were constrained to their final bank values, while the PHQ-9 items were calibrated under the constraints imposed by the anchor items.

Table 5.16.3: IRT Linking Constants

	A	B
Mean/Mean	1.162	0.413
Mean/Sigma	1.234	0.365
Haebara	1.237	0.391
Stocking-Lord	1.222	0.378

The item parameter estimates for the PHQ-9 items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.16.3. The PHQ-9 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-9 as shown in Figure 5.16.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.16.6 displays the differences on the vertical axis.

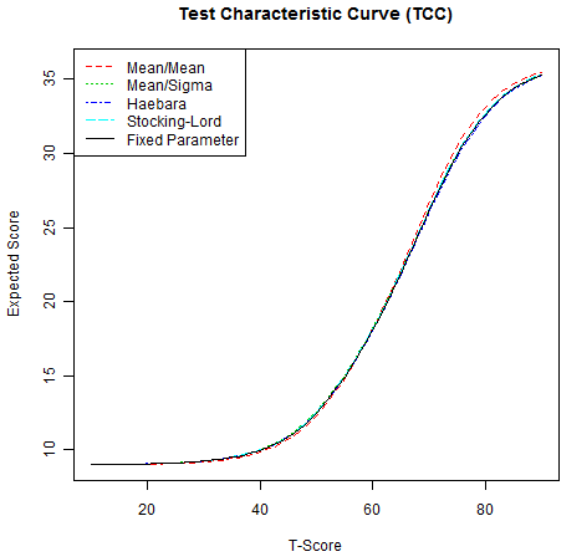

Figure 5.16.5: Test Characteristic Curves (TCC) from Different Linking Methods

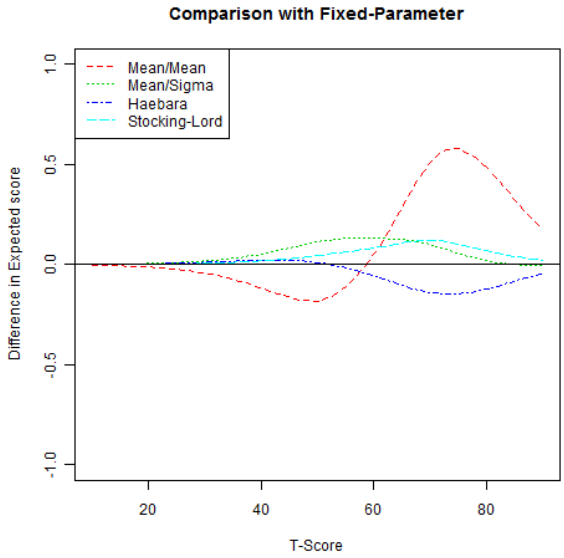

Figure 5.16.6: Difference in Test Characteristic Curves (TCC)

Table 5.16.4 shows the fixed-parameter calibration item parameter estimates for PHQ-9. The marginal reliability estimate for PHQ-9 based on the item parameter estimates was 0.789. The marginal reliability estimates for PROMIS Depression and the combined set were 0.929 and 0.941, respectively. The slope parameter estimates for PHQ-9 ranged from 1.33 to 2.91 with a mean of 1.96. 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 5.16.7 displays the scale information functions for PROMIS Depression, PHQ-9, and the combined set of 29. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.16.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.16.4: Fixed-Parameter Calibration Item Parameter Estimates

Slope	Threshold 1	Threshold 2	Threshold 3
1.953	0.466	1.662	2.266
2.907	0.309	1.423	2.090
1.325	-0.158	1.096	1.994
1.671	-0.401	0.963	1.814
1.481	0.307	1.441	2.256
2.465	0.460	1.414	2.072
1.855	0.811	2.005	2.645
1.815	1.476	2.375	3.113
2.198	1.598	2.441	2.974

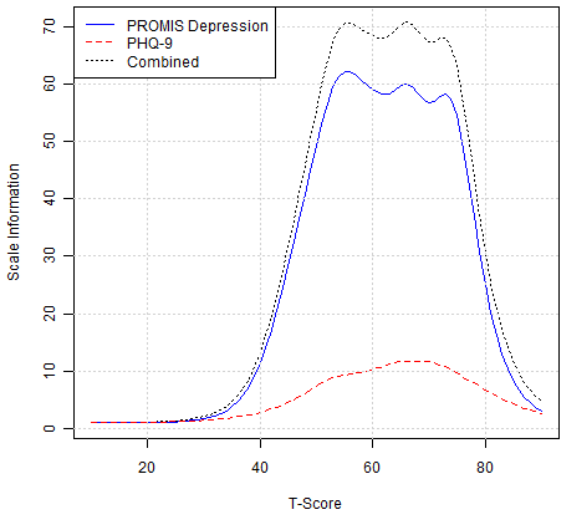

Figure 5.16.7: Comparison of Scale Information Functions

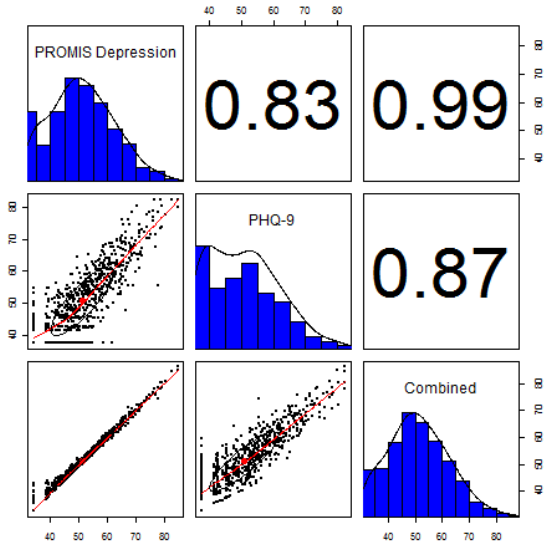

Figure 5.16.8: Comparison of IRT Scaled Scores

5.16.5. 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-9 to a scaled score on PROMIS Depression can be useful. Based on the PHQ-9 item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 46 can be used to map simple raw summed scores from PHQ-9 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.

5.16.6. Equipercentile Linking

We mapped each raw summed score point on PHQ-9 to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on PHQ-9. 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 5.16.9 displays the cumulative distribution functions of the measures. Figure 5.16.10 shows the equipercentile linking functions based on raw summed scores, from PHQ-9 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 47 and Appendix Table 48 show the equipercentile crosswalk tables. The result shown in Appendix Table 47 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 48 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).

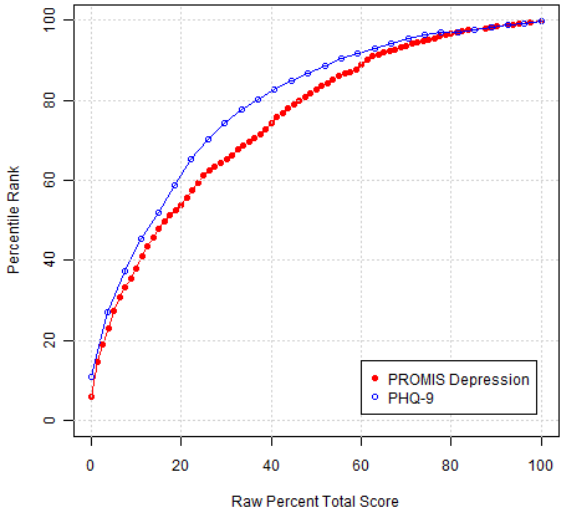

Figure 5.16.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

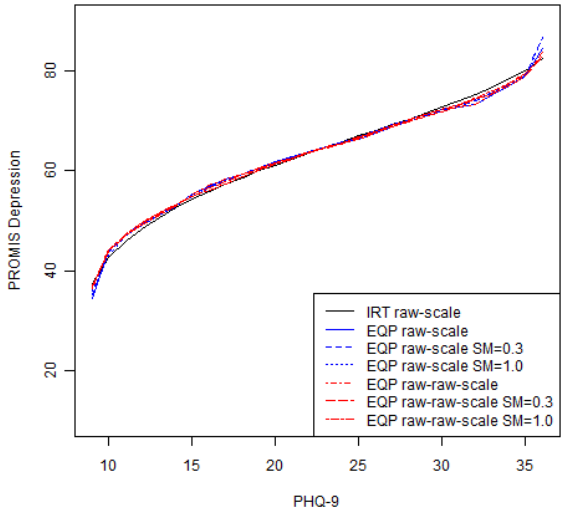

Figure 5.16.10: Equipercentile Linking Functions

5.16.7. 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 5.16.10, 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 5.16.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Depression T-scores and PHQ-9 scores linked to the T-score metric through different methods. In addition to

the seven linking methods previously discussed (see Figure 5.16.10), 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 (0.831), followed by IRT raw-scale (0.807). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (6.342), followed by IRT raw-scale (6.739).

Table 5.16.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.831	0.362	6.336	6.342
IRT raw-scale	0.807	0.434	6.730	6.739
EQP raw-scale SM=0.0	0.794	0.576	7.196	7.214
EQP raw-scale SM=0.3	0.798	0.452	7.086	7.096
EQP raw-scale SM=1.0	0.801	0.287	6.958	6.959
EQP raw-raw-scale SM=0.0	0.801	0.182	6.875	6.873
EQP raw-raw-scale SM=0.3	0.802	0.055	6.805	6.801
EQP raw-raw-scale SM=1.0	0.802	0.081	6.821	6.817

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=748), however, subsetting out a sample was not feasible. Instead, a resampling study was used where 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 5.16.6 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, 0.7. That is, the difference between the mean PROMIS Depression T-score and the mean equated PHQ-9 T-score based on a similar sample of 75 cases is expected to be around ±1.4 (i.e., 2 x 0.70).

Table 5.16.6: 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	0.355	1.249	0.362	0.869	0.359	0.700
IRT raw-scale	0.423	1.323	0.442	0.926	0.435	0.737
EQP raw-scale SM=0.0	0.581	1.413	0.573	0.974	0.570	0.778
EQP raw-scale SM=0.3	0.442	1.390	0.456	0.964	0.468	0.774
EQP raw-scale SM=1.0	0.289	1.366	0.297	0.945	0.288	0.764
EQP raw-raw-scale SM=0.0	0.202	1.366	0.181	0.943	0.183	0.756
EQP raw-raw-scale SM=0.3	0.064	1.339	0.068	0.924	0.057	0.747
EQP raw-raw-scale SM=1.0	0.082	1.349	0.082	0.938	0.088	0.746

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 46) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.17. PROMIS Anxiety and Neuro-QOL Anxiety

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Anxiety, namely the PROMIS Anxiety item bank and Neuro-QOL Anxiety (21 items). The two measures shared 15 common items which served as anchors in linking the Neuro-QOL Anxiety to PROMIS. PROMIS Anxiety was scaled such that higher scores represent higher levels of Anxiety. 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.

5.17.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 75 for PROMIS Anxiety and 105 for Neuro-QOL Anxiety. Figures 5.17.1 and 5.17.2 graphically display the raw summed score distributions of the two measures. Figure 5.17.3 shows the distribution for the combined. Figure 5.17.4 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 Anxiety and Neuro-QOL Anxiety was 0.99. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anxiety and Neuro-QOL Anxiety was 1. The correlations between the combined score and the measures were 0.99 and 1 for PROMIS Anxiety and Neuro-QOL Anxiety, respectively.

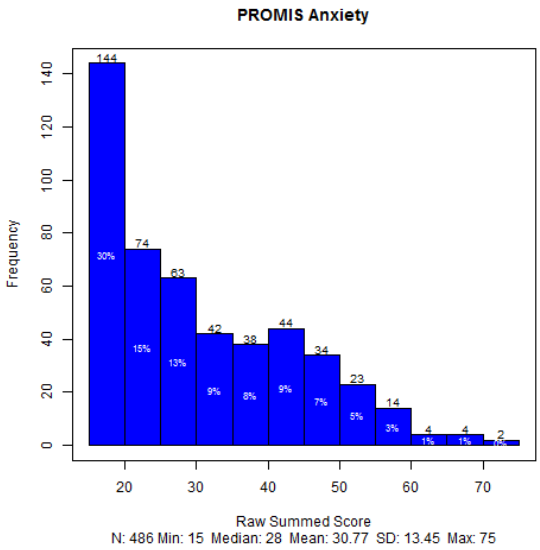

Figure 5.17.1: Raw Summed Score Distribution - PROMIS Anxiety

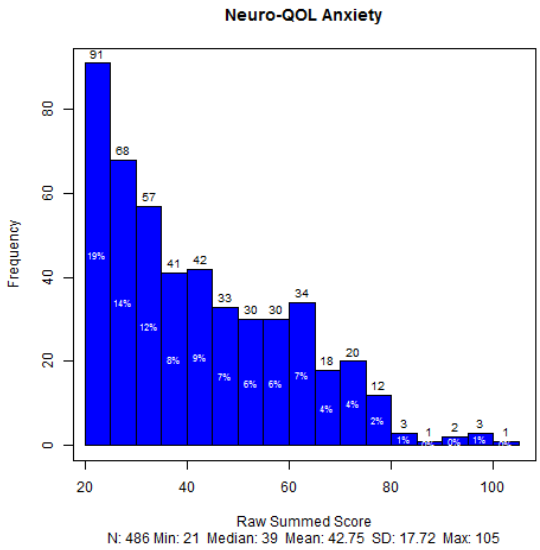

Figure 5.17.2: Raw Summed Score Distribution - Neuro-QOL Anxiety

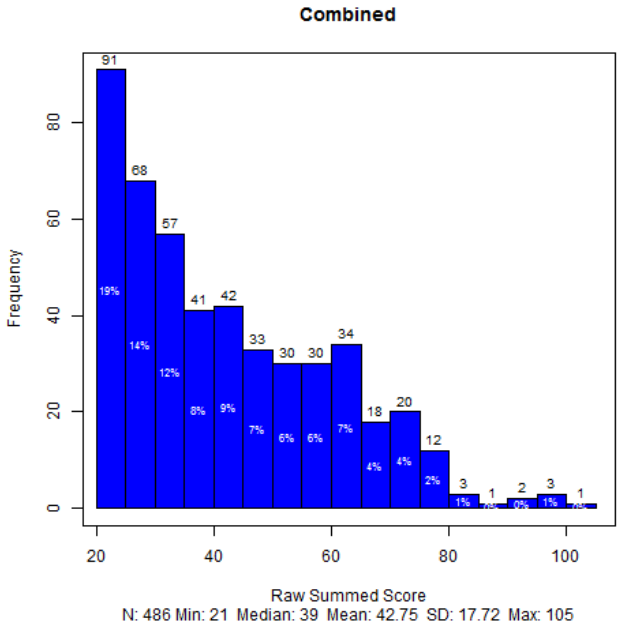

Figure 5.17.3: Raw Summed Score Distribution - Combined

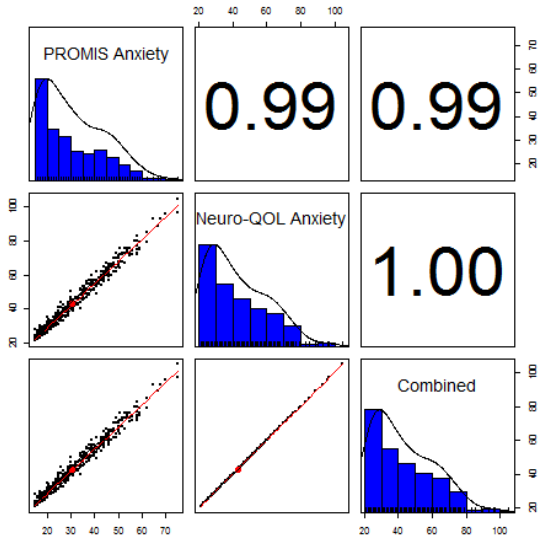

Figure 5.17.4: Scatter Plot Matrix of Raw Summed Scores

5.17.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.17.1 summarizes the results. For PROMIS Anxiety, Cronbach's alpha internal consistency reliability estimate was 0.962 and adjusted (corrected for overlap) item-total correlations ranged from 0.587 to 0.882. For Neuro-QOL Anxiety, alpha was 0.964 and adjusted item-total correlations ranged from 0.547 to 0.879. For the 21 items, alpha was 0.964 and adjusted item-total correlations ranged from 0.547 to 0.879.

Table 5.17.1: Classic Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
PROMS Anxiety	15	0.962	0.587	0.780	0.882
Neuro-QOL Anxiety	21	0.964	0.547	0.738	0.879
Combined	21	0.964	0.547	0.738	0.879

5.17.3. 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 each of the two measures separately and on the combined. Table 5.17.2 summarizes the model fit statistics.

For PROMIS Anxiety, the fit statistics were as follows: CFI = 0.984, TLI = 0.981, and RMSEA = 0.103. For Neuro-QOL Anxiety, CFI = 0.978, TLI = 0.976, and RMSEA = 0.088. For the 21 items, CFI = 0.978, TLI = 0.976, and RMSEA = 0.088. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.17.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Anxiety	15	513	0.984	0.981	0.103
Neuro-QOL Anxiety	21	513	0.978	0.976	0.088
Combined	21	513	0.978	0.976	0.088

5.17.4. Item Response Theory (IRT Linking)

We conducted concurrent calibration on the combined set of 21 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 21 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 15 PROMIS Anxiety items served as anchor items to transform the item parameter estimates for the Neuro-QOL Anxiety items onto the PROMIS Anxiety 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 5.17.3 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 Anxiety items were constrained to their final bank values, while the Neuro-QOL Anxiety items were calibrated, under the constraints imposed by the anchor items.

Table 5.17.3: IRT Linking Constants

	A	B
Mean/Mean	1.163	0.303
Mean/Sigma	1.260	0.224
Haebara	1.245	0.242
Stocking-Lord	1.230	0.241

The item parameter estimates for the Neuro-QOL Anxiety items were linked to the PROMIS Anxiety metric using the transformation constants shown in Table 5.17.3. The Neuro-QOL Anxiety item parameter estimates from the fixed- parameter calibration are considered already on the PROMIS Anxiety metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for Neuro-QOL Anxiety as shown in Figure 5.17.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.17.6 displays the differences on the vertical axis.

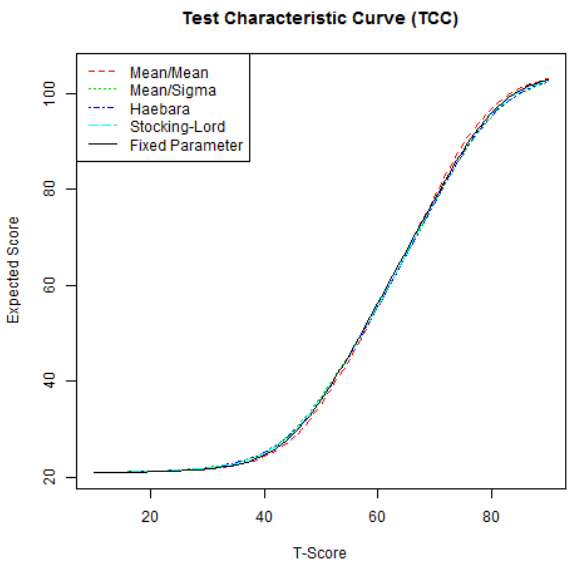

Figure 5.17.5: Test Characteristic Curves (TCC) from Different Linking Methods

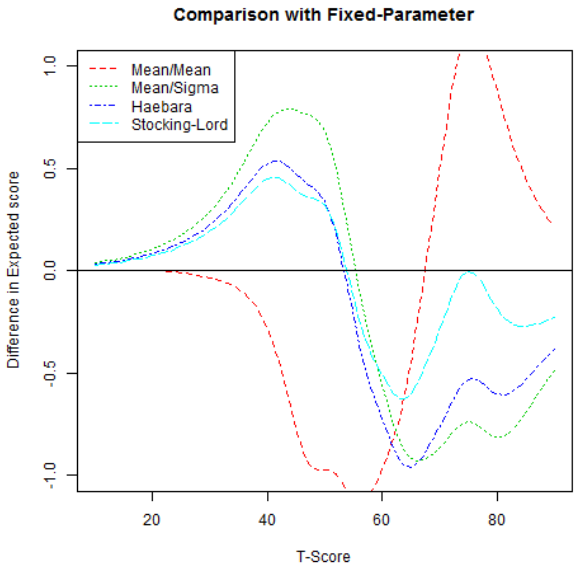

Figure 5.17.6: Difference in Test Charactertistic Curves (TCC)

Table 5.17.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QOL Anxiety. The marginal reliability estimate for Neuro-QOL Anxiety based on the item parameter estimates was 0.938. The marginal reliability estimates for PROMIS Anxiety and the combined set were 0.923 and 0.938, respectively. The slope parameter estimates for Neuro-QOL Anxiety ranged from 1.21 to 3.66 with a mean of 2.58. The slope parameter estimates for PROMIS Anxiety ranged from 1.52 to 3.66 with a mean of 2.91. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.17.7 displays the scale information functions for PROMIS Anxiety, Neuro-QOL Anxiety, and the combined set of 21. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.17.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.17.4: Fixed-Parameter Calibration Item Parameter Estimates for Neuro-QOL Anxiety

a	cb1	cb2	cb3	cb4	NCAT
3.360	-0.190	0.598	1.570	2.450	5
3.550	0.539	1.050	1.870	2.380	5
2.990	0.566	1.300	2.150	3.060	5
1.720	0.010	1.170	2.200	3.110	5
1.950	-0.026	0.879	2.010	3.240	5
2.860	0.436	1.130	2.030	2.780	5
3.030	-0.515	0.316	1.350	2.300	5
3.660	0.364	1.030	1.780	2.620	5
3.400	-0.217	0.632	1.640	2.730	5
3.040	-0.332	0.556	1.460	2.340	5
1.520	-0.830	0.087	1.170	2.380	5
2.410	-0.462	0.398	1.360	2.400	5
3.660	-0.232	0.595	1.560	2.500	5
3.350	-0.509	0.311	1.250	2.300	5
3.130	0.074	0.941	1.850	2.780	5

2.567	-0.101	0.710	1.618	2.518	5
2.016	-0.599	0.409	1.283	2.244	5
1.208	-0.955	0.219	1.537	2.780	5
2.052	1.132	1.810	2.675	3.316	5
1.387	0.814	1.709	3.007	3.673	5
1.373	0.037	1.113	2.064	2.933	5

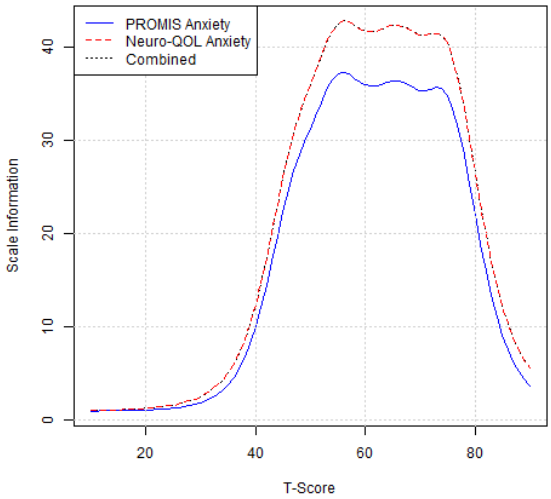

Figure 5.17.7: Comparison of Scale Information Functions

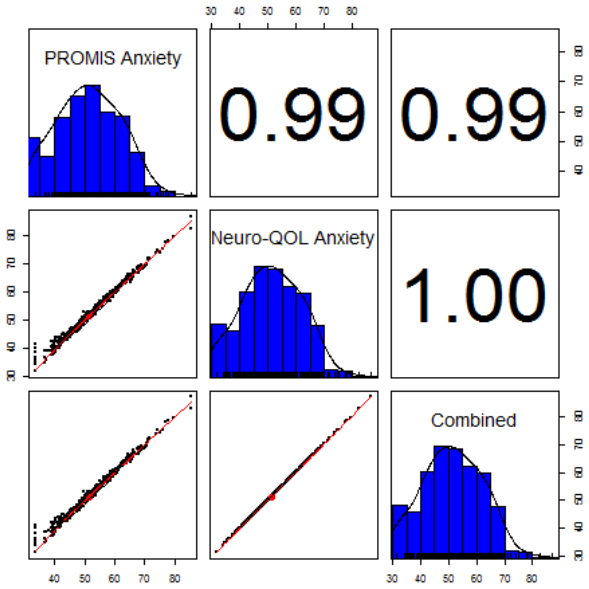

Figure 5.17.8: Comparison of IRT Scaled Scores

5.17.5. 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 Neuro-QOL Anxiety to a scaled score on PROMIS Anxiety can be useful. Based on the Neuro-QOL Anxiety item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 49 can be used to map simple raw summed scores from Neuro-QOL Anxiety to T-score values linked to the PROMIS Anxiety 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.

5.17.6. Equipercentile Linking

We mapped each raw summed score point on Neuro-QOL Anxiety to a corresponding scaled score on PROMIS Anxiety by identifying scores on PROMIS Anxiety that have the same

percentile ranks as scores on Neuro-QOL Anxiety. 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 5.17.9 displays the cumulative distribution functions of the measures. Figure 5.17.10 shows the equipercentile linking functions based on raw summed scores, from Neuro-QOL Anxiety to PROMIS Anxiety. 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 50 and Appendix Table 51 show the equipercentile crosswalk tables. The result shown in Appendix Table 50 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 51 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).

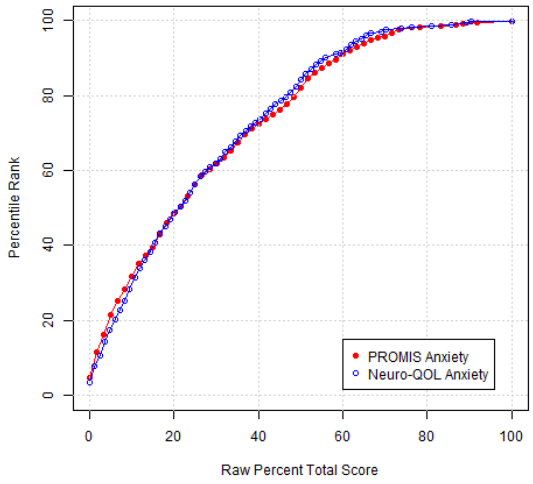

Figure 5.17.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

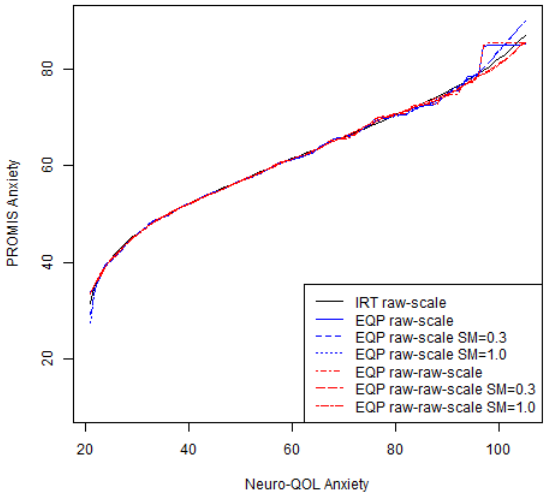

Figure 5.17.10: Equipercentile Linking Functions

5.17.7. 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 over 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 5.17.10, 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 5.17. 8 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Anxiety T-scores and Neuro-QOL Anxiety scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.17.10), 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 (0.994), followed by EQP raw-raw-scale SM=0.0 (0.983). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (1.146), followed by EQP raw-raw-scale SM=1.0 (1.955).

Table 5.17.8: Observed vs Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.994	0.038	1.147	1.146
IRT raw-scale	0.982	0.058	2.023	2.022
EQP raw-scale SM=0.0	0.982	-0.073	1.994	1.993
EQP raw-scale SM=0.3	0.979	0.264	2.261	2.274
EQP raw-scale SM=1.0	0.976	0.418	2.506	2.538
EQP raw-raw-scale SM=0.0	0.983	-0.024	1.959	1.957
EQP raw-raw-scale SM=0.3	0.983	-0.022	1.962	1.960
EQP raw-raw-scale SM=1.0	0.983	-0.048	1.957	1.955

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=486), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=486) over a large number of replications (i.e., 10,000).

Table 5.17.9 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 Anxiety 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, 0.124. That is, the difference between the mean PROMIS Anxiety T-score and the mean equated Neuro-QOL Anxiety T-score based on a similar sample of 75 cases is expected to be around ±0.25(i.e., 2 x 0.124).

Table 5.17.9: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	0.035	0.221	0.037	0.152	0.038	0.124
IRT raw-scale	0.061	0.390	0.063	0.272	0.056	0.216
EQP raw-scale SM=0.0	-0.079	0.384	-0.074	0.267	-0.072	0.210
EQP raw-scale SM=0.3	0.260	0.443	0.264	0.303	0.262	0.240
EQP raw-scale SM=1.0	0.420	0.487	0.412	0.341	0.419	0.264
EQP raw-raw-scale SM=0.0	-0.024	0.382	-0.027	0.265	-0.020	0.206
EQP raw-raw-scale SM=0.3	-0.018	0.379	-0.021	0.263	-0.022	0.208
EQP raw-raw-scale SM=1.0	-0.052	0.384	-0.048	0.264	-0.048	0.210

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 49) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.18. PROMIS Depression and Neuro-QOL Depression

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 and Neuro-QOL Depression (24 items). The two measures shared 18 common items which served as anchors in linking the Neuro-QOL Depression to PROMIS. Depression was scaled such that higher scores represent higher levels of Depression. 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.

5.18.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 90 for PROMIS Depression and 120 for Neuro-QOL Depression. Figures 5.18.1 and 5.18.2 graphically display the raw summed score distributions of the two measures. Figure 5.18.3 shows the distribution for the combined. Figure 5.18.4 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 Neuro-QOL Depression was 1. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and Neuro-QOL Depression was 1. The correlations between the combined score and the measures were 1 and 1 for PROMIS Depression and Neuro-QOL Depression, respectively.

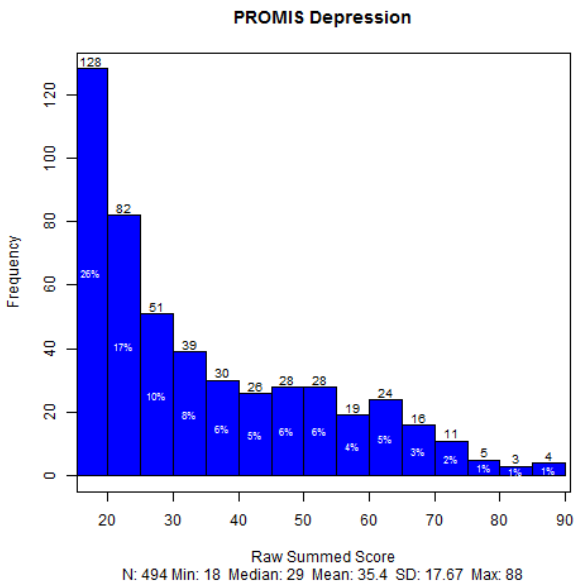

Figure 5.18.1: Raw Summed Score Distribution - PROMIS Depression

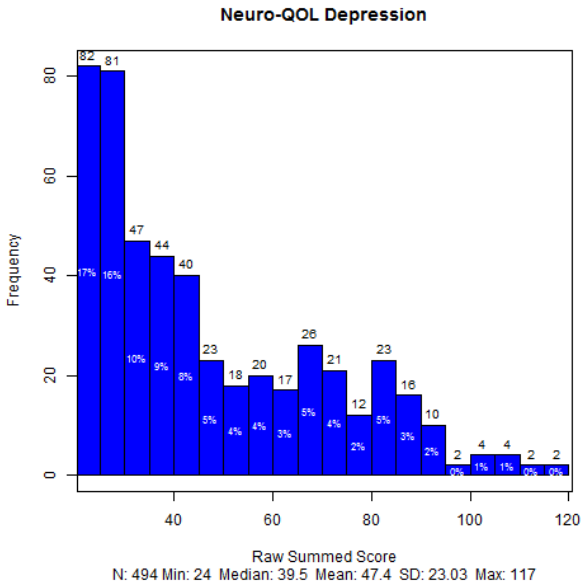

Figure 5.18.2: Raw Summed Score Distribution - Neuro-QOL Depression

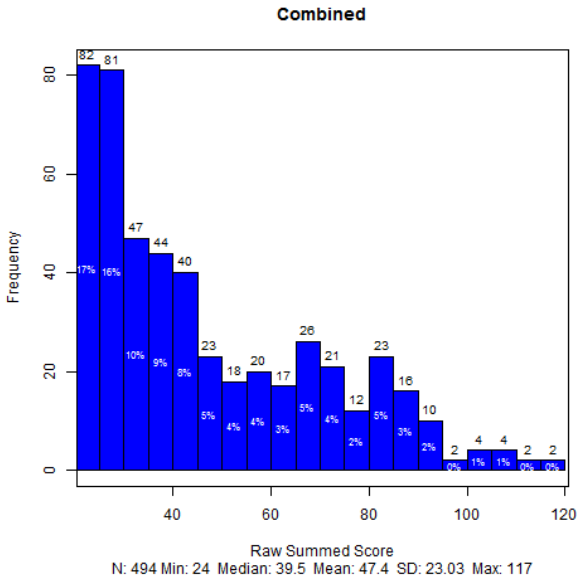

Figure 5.18.3: Raw Summed Score Distribution – Combined

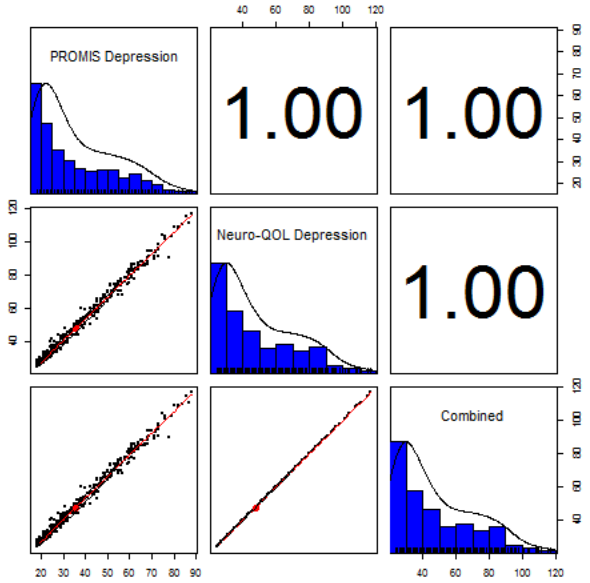

Figure 5.18.4: Scatter Plot Matrix of Raw Summed Scores

5.18.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.18.1 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 0.749 to 0.904. For Neuro-QOL Depression, alpha was 0.981 and adjusted item-total correlations ranged from 0.73 to 0.901. For the 24 items, alpha was 0.981 and adjusted item-total correlations ranged from 0.73 to 0.901.

Table 5.18.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
PROMIS Depression	18	0.979	0.749	0.842	0.904
Neuro-QOL Depression	24	0.981	0.730	0.823	0.901
Combined	24	0.981	0.730	0.823	0.901

5.18.3. 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 each of the two measures separately and on the combined. Table 5.18.2 summarizes the model fit statistics. For PROMIS Depression, the fit statistics were as follows: CFI = 0.993, TLI = 0.993, and

RMSEA = 0.076. For Neuro-QOL Depression, CFI = 0.99, TLI = 0.989, and RMSEA = 0.072. For the 24 items, CFI = 0.99, TLI = 0.989, and RMSEA = 0.072. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.18.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Depression	18	513	0.993	0.993	0.076
Neuro-QOL Depression	24	513	0.990	0.989	0.072
Combined	24	513	0.990	0.989	0.072

5.18.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 24 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 24 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 18 PROMIS Depression items served as anchor items to transform the item parameter estimates for the Neuro-QOL Depression 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 5.18.3 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 Neuro-QOL Depression items were calibrated, under the constraints imposed by the anchor items.

Table 5.18.3: IRT Linking Constants

	A	B
Mean/Mean	1.262	0.373
Mean/Sigma	1.275	0.364
Haebara	1.263	0.380
Stocking-Lord	1.265	0.374

The item parameter estimates for the Neuro-QOL Depression items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.18.3. The Neuro-QOL Depression 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 Neuro-QOL Depression as shown in Figure 5.18.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.18.6 displays the differences on the vertical axis.

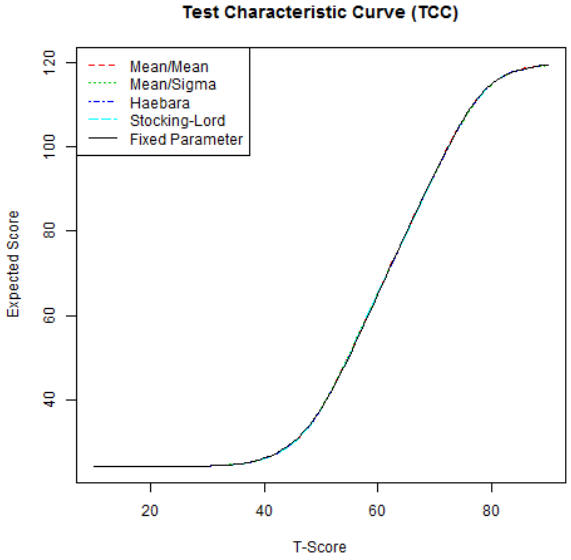

Figure 5.18.5: Test Characteristic Curves (TCC) from Different Linking Methods

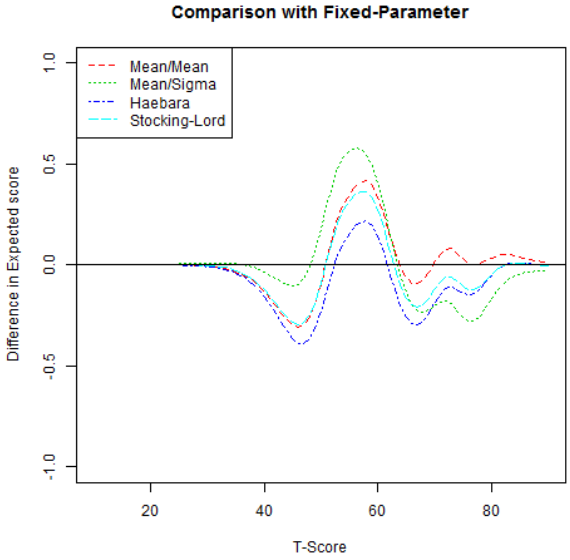

Figure 5.18.6: Difference in Test Characteristic Curves (TCC)

Table 5.18.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QOL Depression. The marginal reliability estimate for Neuro-QOL Depression based on the item parameter estimates was 0.935. The marginal reliability estimates for PROMIS Depression and the combined set were 0.919 and 0.935, respectively. The slope parameter estimates for Neuro-QOL Depression ranged from 1.94 to 4.45 with a mean of 3.13. The slope parameter estimates for PROMIS Depression ranged from 2.38 to 4.45 with a mean of 3.37. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.18.7 displays the scale information functions for PROMIS Depression, Neuro-QOL Depression, and the combined set of 24. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.18.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.18.4: Fixed-Parameter Calibration Item Parameter Estimates for Neuro-QOL Depression

a	cb1	cb2	cb3	cb4	NCAT
4.260	0.401	0.976	1.700	2.440	5
3.930	0.305	0.913	1.590	2.410	5
4.140	0.350	0.915	1.680	2.470	5
2.800	0.148	0.772	1.600	2.540	5
3.660	0.312	0.982	1.780	2.570	5
3.270	-0.498	0.406	1.410	2.380	5
3.240	0.460	1.030	1.830	2.510	5
2.590	-0.079	0.633	1.480	2.330	5
4.340	-0.117	0.598	1.430	2.270	5
3.180	-0.261	0.397	1.310	2.130	5
3.110	0.044	0.722	1.640	2.470	5
3.480	-0.536	0.348	1.350	2.350	5
3.130	0.918	1.480	2.160	2.860	5
4.450	0.558	1.070	1.780	2.530	5
2.830	0.141	0.906	1.850	2.880	5

2.380	-0.458	0.478	1.550	2.630	5
3.190	0.198	0.782	1.530	2.320	5
2.690	-0.299	0.423	1.360	2.310	5
2.128	-0.300	0.485	1.532	2.253	5
2.133	-0.460	0.305	1.480	2.370	5
2.541	0.285	0.928	1.842	2.513	5
1.937	-0.248	0.677	2.004	3.062	5
2.587	0.704	1.220	1.855	2.427	5
3.064	0.266	0.878	1.734	2.361	5

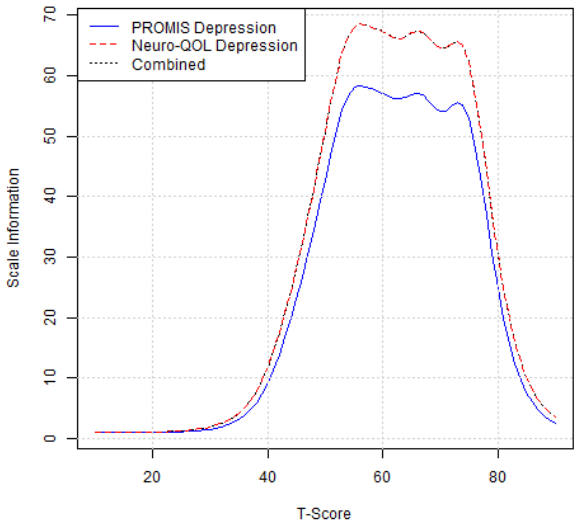

Figure 5.18.7: Comparison of Scale Information Functions

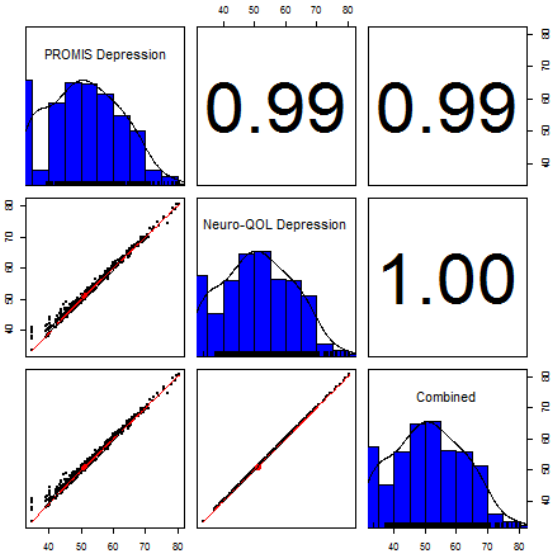

Figure 5.18.8: Comparison of IRT Scaled Scores

5.18.5. 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 Neuro-QOL Depression to a scaled score on PROMIS Depression can be useful. Based on the Neuro-QOL Depression item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 52 can be used to map simple raw summed scores from Neuro-QOL Depression 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.

5.18.6. Equipercentile Linking

We mapped each raw summed score point on Neuro-QOL Depression to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on Neuro-QOL Depression. 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 5.18.9 displays the cumulative distribution functions of the measures. Figure 5.18.10 shows the equipercentile linking functions based on raw summed scores, from Neuro-QOL Depression 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 Tables 53 and 54 show the equipercentile crosswalk tables. The result shown in Appendix Table 53 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 54 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).

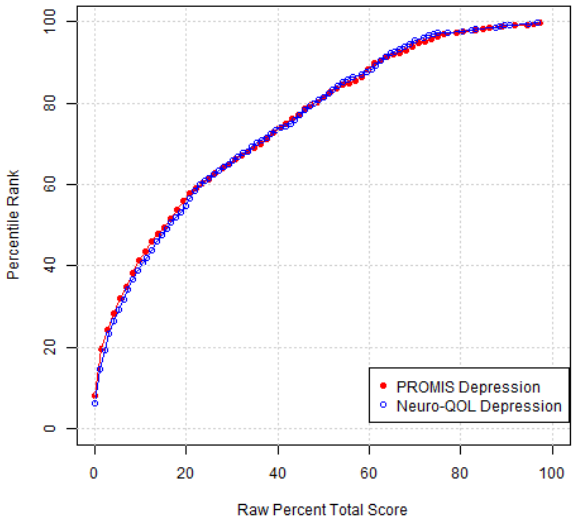

Figure 5.18.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

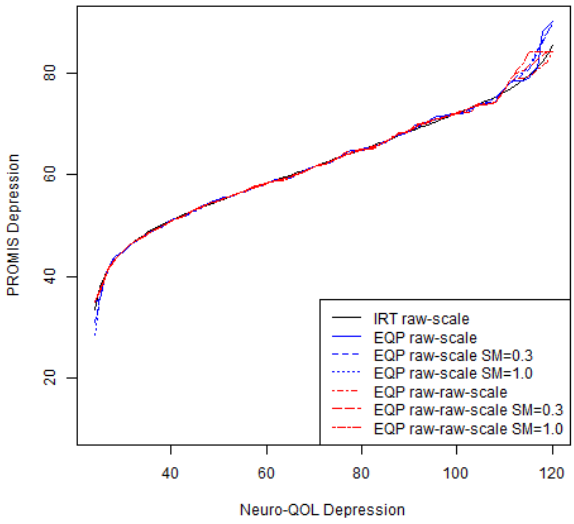

Figure 5.18.9: Equipercentile Linking Functions

5.18.7. 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 5.18.10, 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 5.18.8 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Depression T-scores and Neuro-QOL Depression scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.18.10), 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 (0.995), followed by EQP raw-raw-scale SM=0.3 (0.988). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (1.139), followed by EQP raw-raw-scale SM=0.0 (1.739).

Table 15.8.8: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.995	0.052	1.139	1.139
IRT raw-scale	0.987	0.146	1.801	1.805
EQP raw-scale SM=0.0	0.986	0.081	1.839	1.839
EQP raw-scale SM=0.3	0.983	0.577	2.215	2.287
EQP raw-scale SM=1.0	0.978	0.899	2.755	2.895
EQP raw-raw-scale SM=0.0	0.988	0.048	1.740	1.739
EQP raw-raw-scale SM=0.3	0.988	0.055	1.740	1.739
EQP raw-raw-scale SM=1.0	0.987	0.046	1.768	1.766

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=494), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=494) over a large number of replications (i.e., 10,000).

Table 5.18.9 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, 0.121. That is, the difference between the mean PROMIS Depression T-score and the mean equated Neuro-QOL Depression T-score based on a similar sample of 75 cases is expected to be around ±0.24(i.e., 2 × 0.121).

Table 5.18.9: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	0.053	0.224	0.051	0.152	0.052	0.121
IRT raw-scale	0.146	0.354	0.148	0.244	0.145	0.193
EQP raw-scale SM=0.0	0.074	0.365	0.083	0.247	0.082	0.195
EQP raw-scale SM=0.3	0.575	0.428	0.579	0.295	0.576	0.237
EQP raw-scale SM=1.0	0.906	0.541	0.901	0.371	0.898	0.295
EQP raw-raw-scale SM=0.0	0.049	0.342	0.047	0.236	0.048	0.185
EQP raw-raw-scale SM=0.3	0.052	0.338	0.054	0.233	0.058	0.185
EQP raw-raw-scale SM=1.0	0.048	0.346	0.048	0.235	0.045	0.189

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 52) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.19. PROMIS Physical Function and Neuro-QOL Mobility

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Physical Function, namely the PROMIS Physical Function item bank and Neuro-QOL Mobility (19 items). The two measures shared 8 common items which served as anchors in linking the Neuro-QOL Mobility to PROMIS. PROMIS Physical Function was scaled such that higher scores represent higher levels of Physical Function. 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.

5.19.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 40 for PROMIS Physical Function and 95 for Neuro-QOL Mobility. Figures 5.19.1 and 5.19.2 graphically display the raw summed score distributions of the two measures. Figure 5.19.3 shows the distribution for the combined. Figure 5.19.4 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 Physical Function and Neuro-QOL Mobility was 0.97. The disattenuated (corrected for unreliabilities) correlation between PROMIS Physical Function and Neuro-QOL Mobility was 1. The correlations between the combined score and the measures were 0.97 and 1 for PROMIS Physical Function and Neuro-QOL Mobility, respectively.

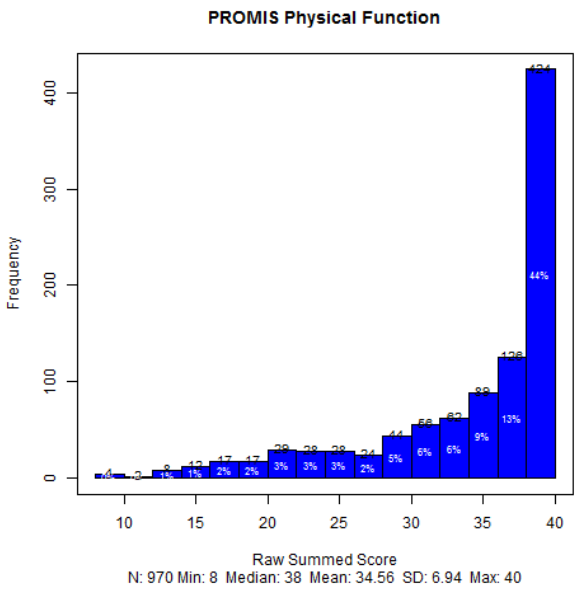

Figure 5.19.1: Raw Summed Score Distribution - PROMIS Physical Function

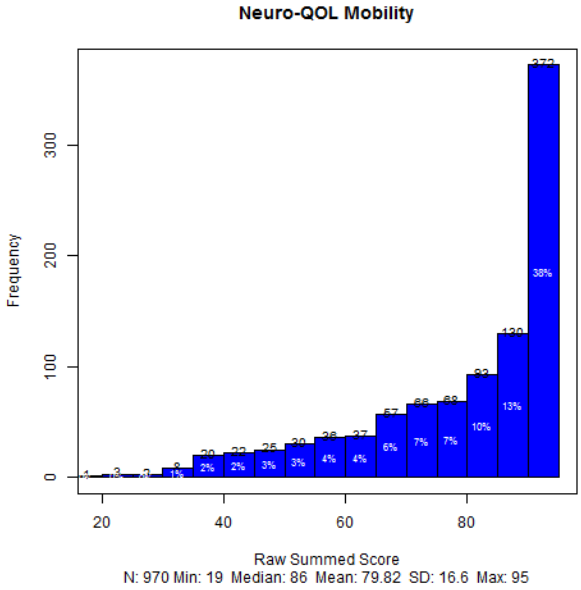

Figure 5.19.2: Raw Summed Score Distribution - Neuro-QOL Mobility

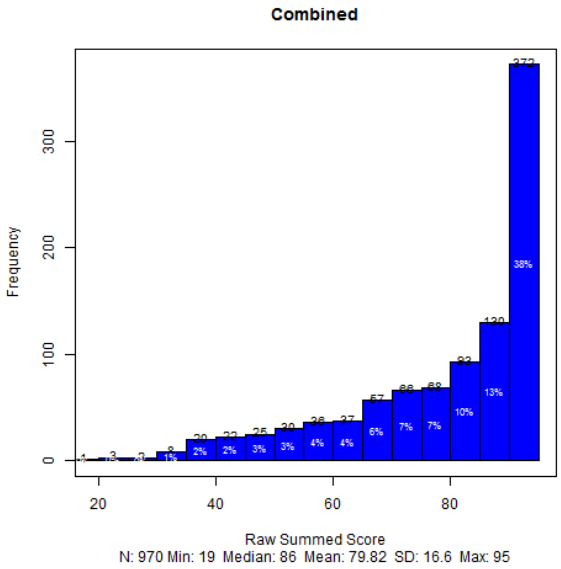

Figure 5.19.3: Raw Summed Score Distribution – Combined

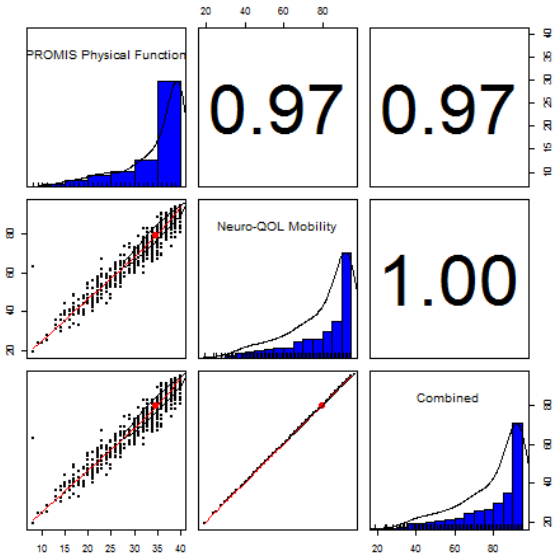

Figure 5.19.4: Scatter Plot Matrix of Raw Summed Scores

5.19.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.19.1 summarizes the results. For PROMIS Physical Function, Cronbach's alpha internal consistency reliability estimate was 0.926 and adjusted (corrected for overlap) item-total correlations ranged from 0.743 to 0.809. For Neuro-QOL Mobility, alpha was 0.967 and adjusted item-total correlations ranged from 0.675 to 0.844. For the 19 items, alpha was 0.967 and adjusted item-total correlations ranged from 0.675 to 0.844.

Table 5.9.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
PROMIS Physical Function	8	0.926	0.743	0.770	0.809
Neuro-QOL Mobility	19	0.967	0.675	0.780	0.844
Combined	19	0.967	0.675	0.780	0.844

5.19.3. 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 each of the two measures separately and on the combined. Table 5.19.2 summarizes the model fit statistics. For PROMIS Physical Function, the fit statistics were as follows: CFI = 0.995, TLI = 0.993, and

RMSEA = 0.07. For Neuro-QOL Mobility, CFI = 0.977, TLI = 0.974, and RMSEA = 0.1. For the 19 items, CFI = 0.977, TLI = 0.974, and RMSEA = 0.1. The main interest of the current analysis is whether the combined measure is essential unidiemnsional.

Table 5.19.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Physical Function	8	1041	0.995	0.993	0.070
Neuro-QOL Mobility	19	1044	0.977	0.974	0.100
Combined	19	1044	0.977	0.974	0.100

5.19.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 19 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 19 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 8 PROMIS Physical Function items served as anchor items to transform the item parameter estimates for the Neuro-QOL Mobility items onto the PROMIS Physical Function 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 5.19.3 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 Physical Function items were constrained to their final bank values, while the Neuro-QOL Mobility items were calibrated, under the constraints imposed by the anchor items.

Table 5.19.3: IRT Linking Constants

	A	B
Mean/Mean	0.946	-0.836
Mean/Sigma	1.132	-0.628
Haebara	1.098	-0.677
Stocking-Lord	1.077	-0.681

The item parameter estimates for the Neuro-QOL Mobility items were linked to the PROMIS Physical Function metric using the transformation constants shown in Table 5.19.3. The Neuro-QOL Mobility item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Physical Function metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for Neuro-QOL Mobility as shown in Figure 5.19.5. Using thefixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.19.6 displays the differences on the vertical axis.

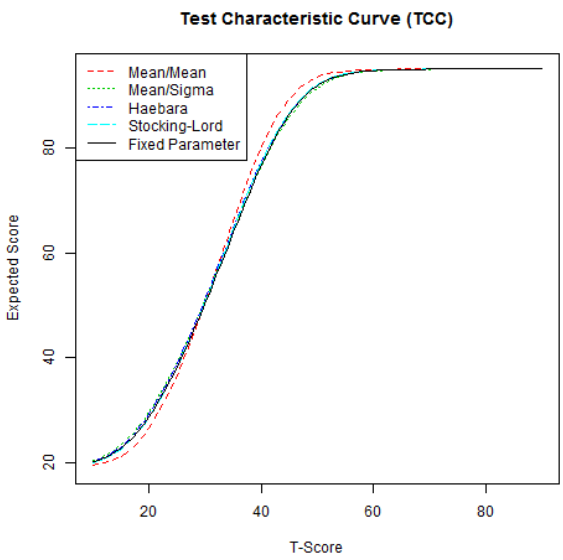

Figure 5.19.5: Test Characteristic Curves (TCC) from Different Linking Methods

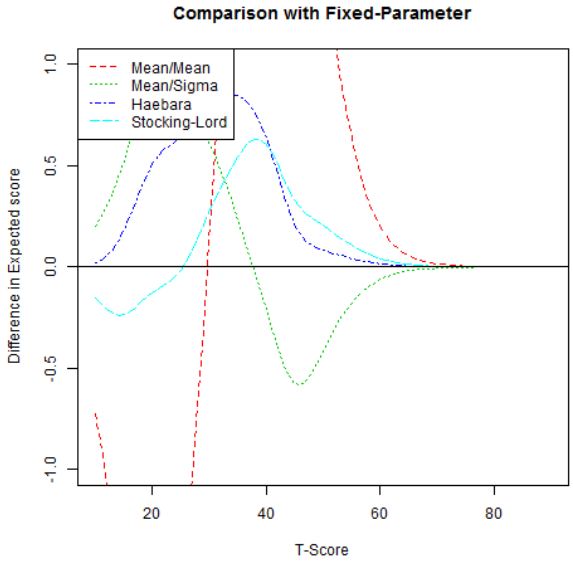

Figure 5.19.6: Difference in Test Characteristic Curves (TCC)

Table 5.19.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QOL Mobility. The marginal reliability estimate for Neuro-QOL Mobility based on the item parameter estimates was 0.836. The marginal reliability estimates for PROMIS Physical Function and the combined set were 0.666 and 0.836, respectively. The slope parameter estimates for Neuro-QOL Mobility ranged from 2.24 to 4.44 with a mean of 3.32. The slope parameter estimates for PROMIS Physical Function ranged from 3.11 to 4.44 with a mean of 3.65. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.19.7 displays the scale information functions for PROMIS Physical Function, Neuro-QOL Mobility, and the combined set of 19. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.19.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.19.4: Fixed-Parameter Calibration Item Parameter Estimates for Neuro-QOL Mobility

a	cb1	cb2	cb3	cb4	NCAT
3.460	-2.700	-2.040	-1.520	-0.750	5
4.290	-1.910	-1.580	-1.190	-0.681	5
4.070	-2.860	-2.270	-1.730	-1.030	5
3.260	-2.130	-1.550	-1.050	-0.306	5
3.390	-3.380	-2.760	-2.200	-1.490	5
4.440	-2.550	-1.970	-1.450	-0.819	5
3.200	-3.570	-2.680	-1.940	-1.070	5
3.110	-3.110	-2.780	-2.210	-1.460	5
3.531	-2.788	-2.153	-1.463	-0.658	5
3.020	-3.544	-2.773	-1.941	-1.110	5
2.415	-3.553	-2.690	-2.037	-1.326	5
3.328	-2.666	-1.653	-1.100	-0.191	5

3.165	-2.735	-2.017	-1.428	-0.727	5
3.629	-2.869	-1.962	-1.248	-0.526	5
3.598	-3.238	-2.741	-2.165	-1.460	5
2.616	-1.701	-1.300	-0.888	-0.265	5
2.712	-2.102	-1.291	-0.694	0.251	5
3.566	-1.811	-1.275	-0.833	-0.233	5
2.239	-2.790	-1.966	-1.411	-0.695	5

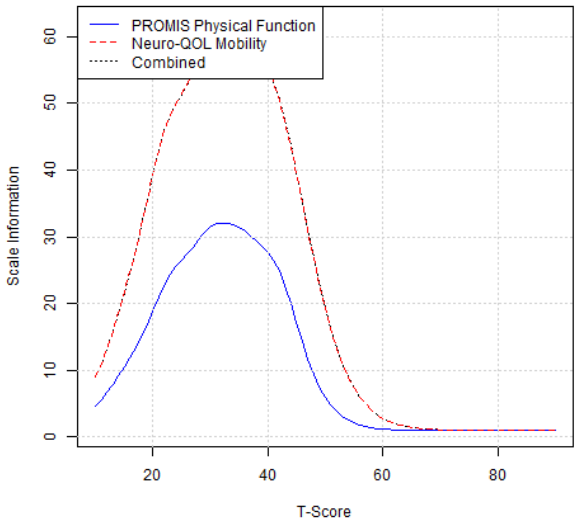

Figure 5.19.7: Comparison of Scale Information Functions

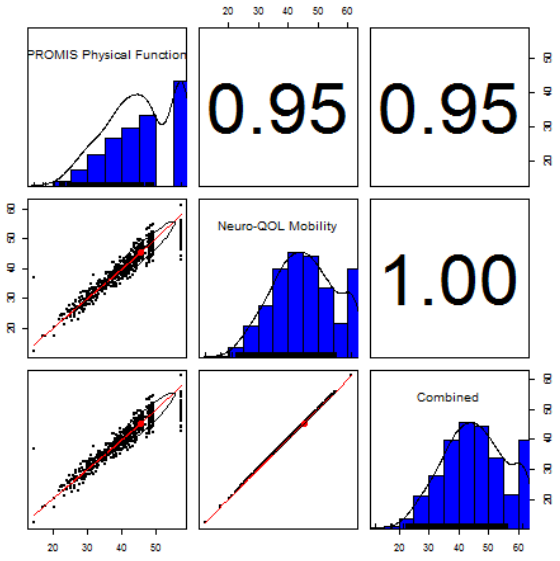

Figure 5.19.8: Comparison of IRT Scaled Scores

5.19.5. 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 Neuro-QOL Mobility to a scaled score on PROMIS Physical Function can be useful. Based on the Neuro-QOL Mobility item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 55 can be used to map simple raw summed scores from Neuro-QOL Mobility to T-score values linked to the PROMIS Physical Function 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.

5.19.6. Equipercentile Linking

We mapped each raw summed score point on Neuro-QOL Mobility to a corresponding scaled score on PROMIS Physical Function by identifying scores on PROMIS Physical Function that have the same percentile ranks as scores on Neuro-QOL Mobility. 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 5.19.9 displays the cumulative distribution functions of the measures. Figure 5.19.10 shows the equipercentile linking functions based on raw summed scores from Neuro-QOL Mobility to PROMIS Physical Function. 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 Tables 56 and 57 show the equipercentile crosswalk tables. The result shown in Appendix Table 56 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 57 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).

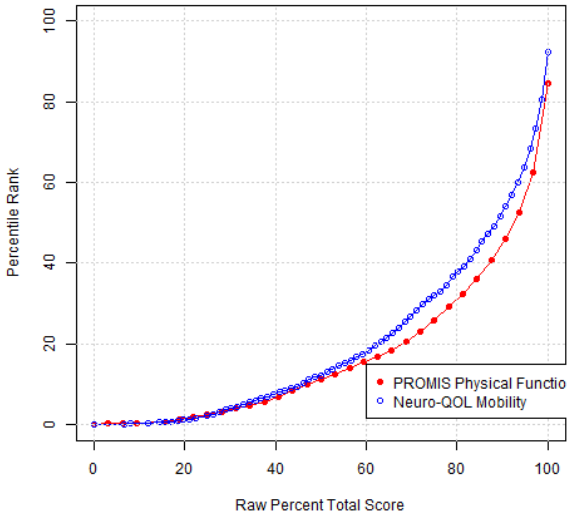

Figure 5.19.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

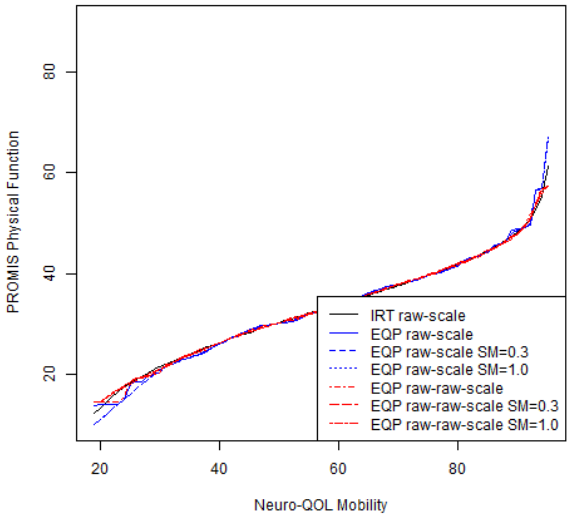

Figure 5.19.10: Equipercentile Linking Functions

5.19.7. 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 5.19.10, 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 5.19.8 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Physical Function T-scores and Neuro-QOL Mobility scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.19.10), 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, EQP raw-raw-scale SM=0.0 produced the best result (0.956), followed by EQP raw-raw-scale SM=0.3 (0.955). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). EQP raw-raw-scale SM=0.0 yielded smallest RMSD (2.965), followed by EQP raw-raw-scale SM=0.3 (3.015).

Table 5.19.8: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.948	0.049	3.301	3.300
IRT raw-scale	0.944	0.006	3.420	3.418
EQP raw-scale SM=0.0	0.951	0.295	3.106	3.119
EQP raw-scale SM=0.3	0.928	-1.043	4.530	4.646
EQP raw-scale SM=1.0	0.924	-1.200	4.787	4.933
EQP raw-raw-scale	0.956	0.466	2.929	2.965
EQP raw-raw-scale	0.955	0.507	2.973	3.015
EQP raw-raw-scale	0.953	0.496	3.018	3.057

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=970), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=970) over a large number of replications (i.e., 10,000).

Table 5.19.9 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 Physical Function 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, EQP raw-raw-scale SM=0.0 produced the smallest standard error, 0.324. That is, the difference between the mean PROMIS Physical Function T-score and the mean equated Neuro-QOL Mobility T-score based on a similar sample of 75 cases is expected to be around ±0.65(i.e., 2 × 0.324).

Table 5.19.9: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	0.043	0.654	0.046	0.462	0.044	0.365
IRT raw-scale	0.001	0.676	-0.003	0.470	0.005	0.378
EQP raw-scale SM=0.0	0.282	0.607	0.303	0.434	0.299	0.346
EQP raw-scale SM=0.3	-1.042	0.891	-1.048	0.626	-1.040	0.506
EQP raw-scale SM=1.0	-1.212	0.940	-1.199	0.659	-1.203	0.534
EQP raw-raw-scale SM=0.0	0.465	0.577	0.466	0.410	0.464	0.324
EQP raw-raw-scale SM=0.3	0.504	0.587	0.507	0.412	0.510	0.328
EQP raw-raw-scale SM=1.0	0.491	0.597	0.492	0.416	0.496	0.335

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 55) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.20. PROMIS Physical Function and Neuro-QOL Upper Extremity

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Physical Function, namely the PROMIS Physical Function item bank and Neuro-QOL Upper Extremity (20 items). The two measures shared 11 common items which served as anchors in linking the Neuro-QOL Upper Extremity to PROMIS. PROMIS Physical Function was scaled such that higher scores represent higher levels of Physical Function. 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.

5.20.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 55 for PROMIS Physical Function and 100 for Neuro-QOL Upper Extremity. Figures 5.20.1 and 5.20.2 graphically display the raw summed score distributions of the two measures. Figure 5.20.3 shows the distribution for the combined. Figure 5.20.4 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 Physical Function and Neuro-QOL Upper Extremity was 0.99. The disattenuated (corrected for unreliabilities) correlation between PROMIS Physical Function and Neuro-QOL Upper Extremity was 1. The correlations between the combined score and the measures were 0.99 and 1 for PROMIS Physical Function and Neuro-QOL Upper Extremity, respectively.

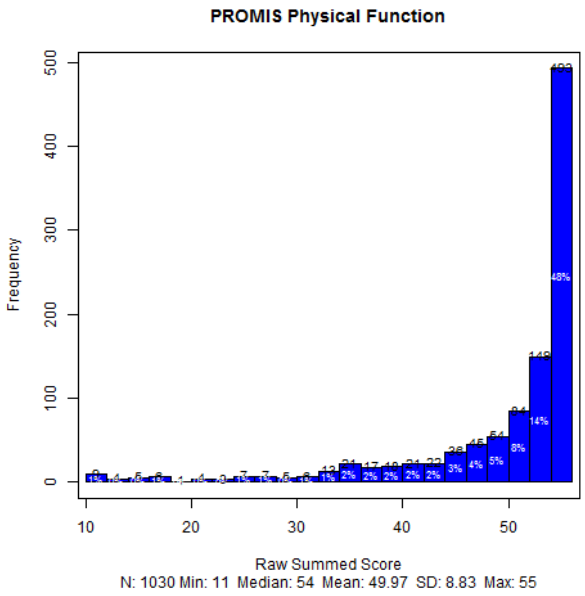

Figure 5.20.1: Raw Summed Score Distribution – PROMIS Physical Function

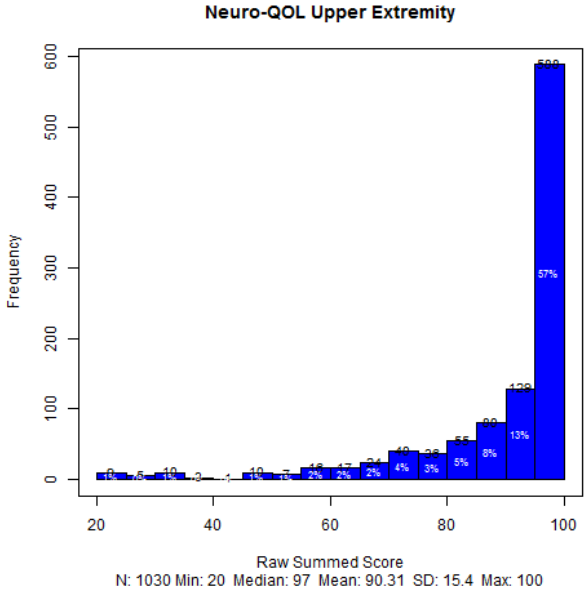

Figure 5.20.2: Raw Summed Score Distribution – Neuro-QOL Upper Extremity

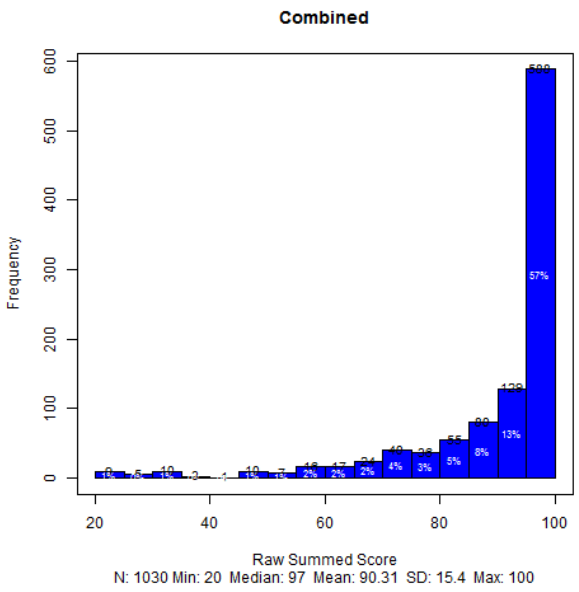

Figure 5.20.3: Raw Summed Score Distribution – Combined

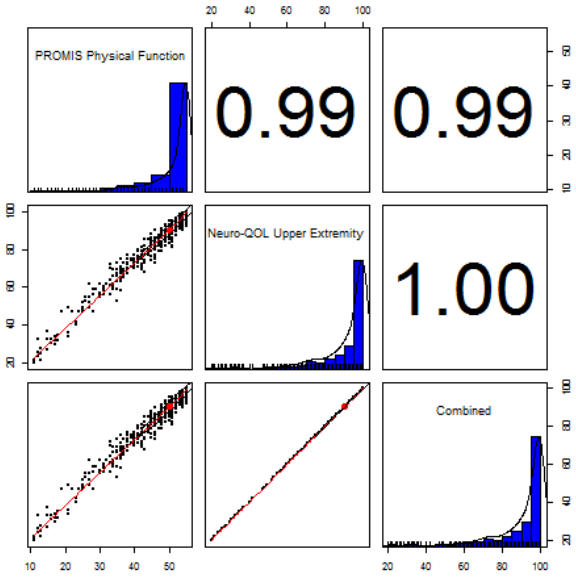

Figure 5.20.4: Scatter Plot Matrix of Raw Summed Scores

5.20.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.20.1 summarizes the results. For PROMIS Physical Function, Cronbach's alpha internal consistency reliability estimate was 0.955 and adjusted (corrected for overlap) item-total correlations ranged from 0.71 to 0.844. For Neuro-QOL Upper Extremity, alpha was 0.969 and adjusted item-total correlations ranged from 0.667 to 0.848. For the 20 items, alpha was 0.969 and adjusted item-total correlations ranged from 0.667 to 0.848.

Table 5.20.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
PROMIS Physical Function	11	0.955	0.710	0.801	0.844
Neuro-QOL Upper Extremity	20	0.96	0.66	0.784	0.84
Combined	20	0.96	0.66	0.784	0.84

5.20.3. 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 each of the two measures separately and on the combined. Table 5.20.2 summarizes the model fit statistics. For PROMIS Physical Function, the fit statistics were as follows: CFI = 0.989, TLI = 0.986, and RMSEA = 0.088. For Neuro-QOL Upper Extremity, CFI = 0.978, TLI = 0.976, and RMSEA = 0.084. For the 20 items, CFI = 0.978, TLI = 0.976, and RMSEA = 0.084. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.20.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Physical Function	11	1092	0.989	0.986	0.088
Neuro-QOL Upper Extremity	20	1093	0.978	0.976	0.084
Combined	20	1093	0.978	0.976	0.084

5.20.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 20 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 20 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 11 PROMIS Physical Function items served as anchor items to transform the item parameter estimates for the Neuro-QOL Upper Extremity items onto the PROMIS Physical Function 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 5.20.3 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 Physical Function items were constrained to their final bank values, while the Neuro-QOL Upper Extremity items were calibrated, under the constraints imposed by the anchor items.

Table 5.20.3: IRT Linking Constants

	A	B
Mean/Mean	1.774	-0.942
Mean/Sigma	1.665	-1.048
Haebara	1.664	-1.041
Stocking-Lord	1.740	-0.982

The item parameter estimates for the Neuro-QOL Upper Extremity items were linked to the PROMIS Physical Function metric using the transformation constants shown in Table 5.20.3. The Neuro-QOL Upper Extremity item parameter estimates from the fixed-parameter calibration

are considered already on the PROMIS Physical Function metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for Neuro-QOL Upper Extremity as shown in Figure 5.20.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.20.6 displays the differences on the vertical axis.

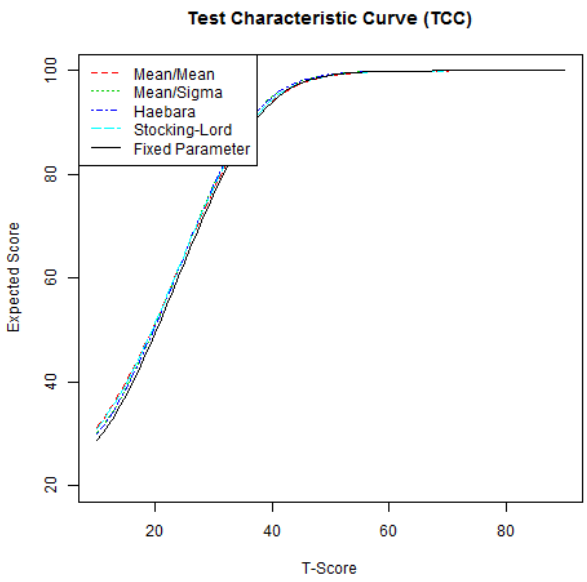

Figure 5.20.5: Test Characteristic Curves (TCC) from Different Linking Methods

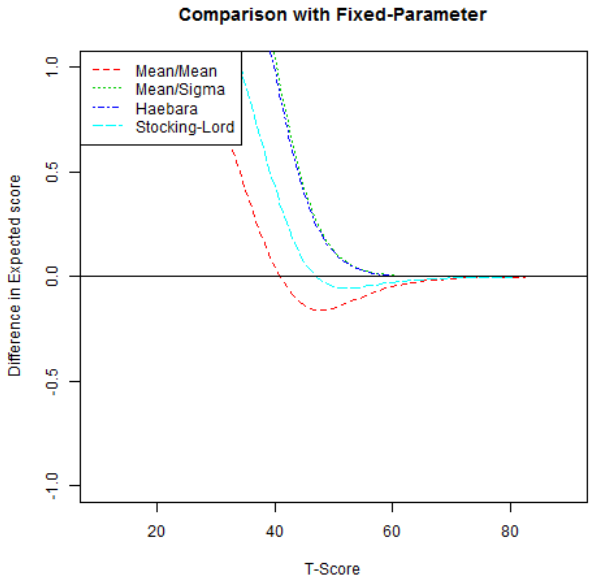

Figure 5.20.6: Difference in Test Characteristic Curves (TCC)

Table 5.20.4: Fixed-Parameter Estimates for Neuro-QOL Upper Extremityshows the fixed-parameter calibration item parameter estimates for Neuro-QOL Upper Extremity. The marginal reliability estimate for Neuro-QOL Upper Extremity based on the item parameter estimates was 0.664. The marginal reliability estimates for PROMIS Physical Function and the combined set were 0.513 and 0.664, respectively. The slope parameter estimates for Neuro-QOL Upper Extremity ranged from 1.73 to 3.58 with a mean of 2.59. The slope parameter estimates for PROMIS Physical Function ranged from 2.09 to 3.58 with a mean of 2.74. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.20.7 displays the scale information functions for PROMIS Physical Function, Neuro-QOL Upper Extremity, and the combined set of 20. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.20.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.20.4: Fixed-Parameter Estimates for Neuro-QOL Upper Extremity

a	cb1	cb2	cb3	cb4	NCAT
3.090	-2.920	-2.210	-1.630	-0.898	5
2.360	-3.770	-3.080	-2.470	-1.850	5
2.090	-3.940	-3.510	-2.660	-1.950	5
3.310	-3.180	-2.730	-2.080	-1.350	5

2.730	-3.780	-3.320	-2.860	-2.350	5
2.150	-3.870	-3.290	-2.620	-1.920	5
3.580	-3.350	-2.650	-2.070	-1.480	5
2.320	-3.870	-3.010	-2.390	-1.690	5
3.110	-3.620	-2.780	-2.290	-1.600	5
3.320	-3.170	-2.880	-2.340	-1.760	5
2.090	-3.660	-3.170	-2.620	-2.020	5
2.098	-4.392	-3.906	-3.120	-2.338	5
2.184	-4.203	-3.576	-2.721	-2.178	5
3.394	-3.491	-2.850	-2.263	-1.504	5
3.127	-3.499	-2.817	-2.164	-1.381	5
2.574	-3.549	-2.501	-1.830	-1.084	5
2.038	-3.448	-2.590	-1.956	-1.035	5
2.215	-2.219	-1.865	-1.323	-0.579	5
2.358	-3.693	-3.131	-2.447	-1.607	5
1.731	-4.509	-2.991	-2.255	-1.570	5

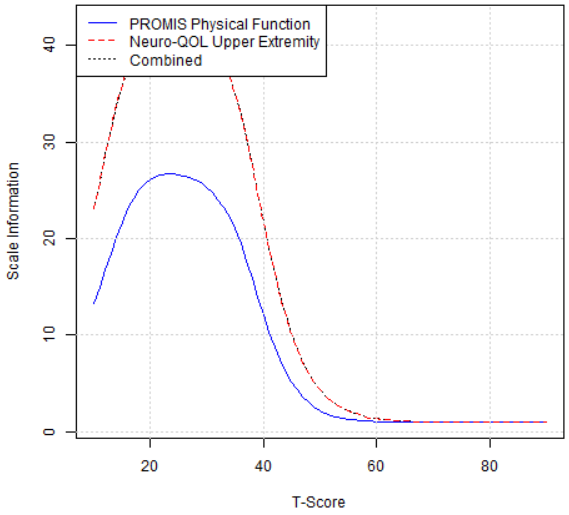

Figure 5.20.7: Comparison of Scale Information Functions

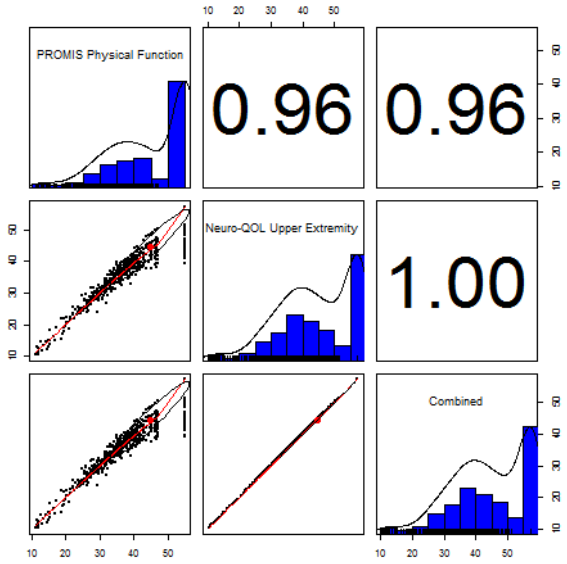

Figure 5.20.8: Comparison of IRT Scaled Scores

5.20.5. 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 Neuro-QOL Upper Extremity to a scaled score on PROMIS Physical Function can be useful. Based on the Neuro-QOL Upper Extremity item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 58 can be used to map simple raw summed scores from Neuro-QOL Upper Extremity to T-score values linked to the PROMIS Physical Function 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.

5.20.6. Equipercentile Linking

We mapped each raw summed score point on Neuro-QOL Upper Extremity to a corresponding scaled score on PROMIS Physical Function by identifying scores on PROMIS Physical Function that have the same percentile ranks as scores on Neuro-QOL Upper Extremity. 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 5.20.9 displays the cumulative distribution functions of the measures. Figure 5.20.10 shows the equipercentile linking functions based on raw summed scores, from Neuro-QOL Upper Extremity to PROMIS Physical Function. 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 Tables 59 and 60 show the equipercentile crosswalk tables. The result shown in Appendix Table 59 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 60 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).

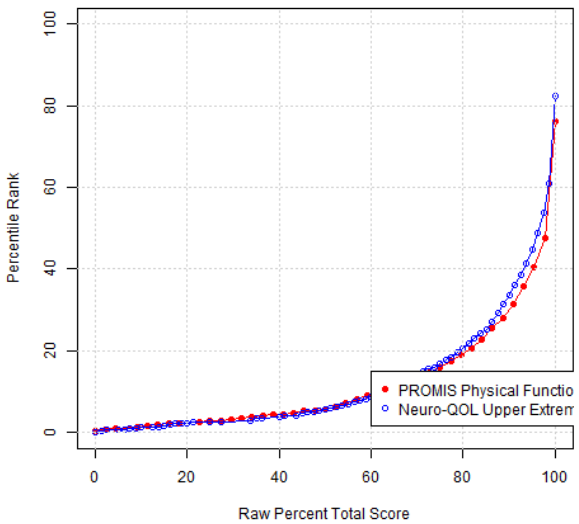

Figure 5.20.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

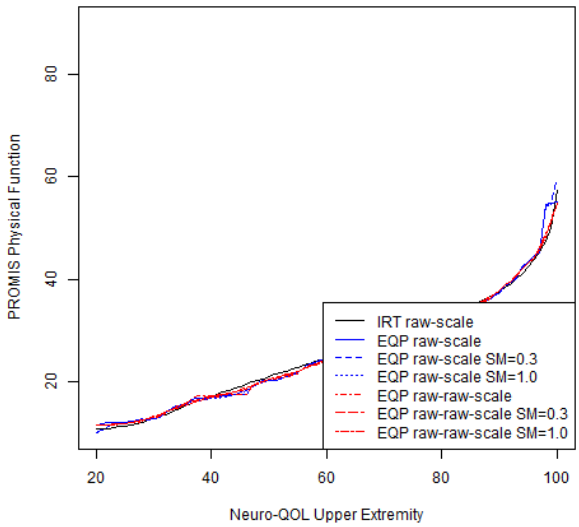

Figure 5.20.10: Equipercentile Linking Functions

5.20.7. 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 5.20.10, 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 5.20.8 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Physical Function T-scores and Neuro-QOL Upper Extremity scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 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, EQP raw-raw-scale SM=0.3 produced the best result (0.959), followed by EQP raw-raw-scale SM=1.0 (0.959). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). EQP raw-raw-scale SM=0.3 yielded smallest RMSD (3.336), followed by EQP raw-raw-scale SM=0.0 (3.353).

Table 5.20.8: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.957	0.062	3.450	3.449
IRT raw-scale	0.950	0.051	3.766	3.765
EQP raw-scale SM=0.0	0.954	0.035	3.550	3.548
EQP raw-scale SM=0.3	0.952	-1.533	4.225	4.492
EQP raw-scale SM=1.0	0.952	-1.538	4.218	4.488
EQP raw-raw-scale SM=0.0	0.959	0.711	3.278	3.353
EQP raw-raw-scale SM=0.3	0.959	0.713	3.261	3.336
EQP raw-raw-scale SM=1.0	0.959	0.747	3.272	3.354

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=1030), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1030) over a large number of replications (i.e., 10,000).

Table 5.20.9 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 Physical Function 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, EQP raw-raw-scale SM=0.3 produced the smallest standard error, 0.36. That is, the difference between the mean PROMIS Physical Function T-score and the mean equated Neuro-QOL Upper Extremity T-score based on a similar sample of 75 cases is expected to be around ±0.72(i.e., 2×0.36).

Table 5.20.9: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	0.054	0.678	0.067	0.474	0.062	0.383
IRT raw-scale	0.057	0.739	0.049	0.518	0.048	0.422
EQP raw-scale SM=0.0	0.050	0.699	0.023	0.492	0.041	0.402
EQP raw-scale SM=0.3	-1.512	0.842	-1.533	0.583	-1.528	0.468
EQP raw-scale SM=1.0	-1.536	0.835	-1.539	0.583	-1.533	0.466
EQP raw-raw-scale SM=0.0	0.718	0.647	0.718	0.457	0.708	0.362
EQP raw-raw-scale SM=0.3	0.718	0.644	0.714	0.457	0.710	0.360
EQP raw-raw-scale SM=1.0	0.738	0.651	0.750	0.451	0.745	0.362

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 58) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

6. References

American Psychiatric Association. Task Force on DSM-IV. (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision*. Washington, D.C.: American Psychiatric Association.

Berndt, E., Kallich, J., McDermott, A., Xu, X., Lee, H., & Glaspy, J. (2005). Reductions in anaemia and fatigue are associated with improvements in productivity in cancer patients receiving chemotherapy. *PharmacoEconomics, 23*(5), 505-514.

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

Brodsky, R.A., Young, N.S., Antonioli, E., Risitano, A.M., Schrezenmeier, H., Schubert, J., . . . Hillmen, P. (2008) Multicenter phase 3 study of the complement inhibitor eculizumab for the treatment of patients with paroxysmal nocturnal hemoglobinuria. *Blood. 111*(4), 1840-1847.

Brucker, P. S., Yost, K., Cashy, J., Webster, K., & Cella, D. (2005). General population and cancer patient norms for the Functional Assessment of Cancer Therapy-General (FACT-G). *Evaluation & the Health Professions, 28*(2), 192-211.

Buss, A.H, & Perry, M.P. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology 63*, 452-459.

Cella, D., Lai, J. S., Chang, C. H., Peterman, A., & Slavin, M. (2002). Fatigue in cancer patients compared with fatigue in the general United States population. *Cancer, 94*(2), 528-538.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group During its First Two Years. *Medical Care, 45*(5 Suppl 1), S3-S11.

Cella, D., Yount, S., Sorensen, M., Chartash, E., Sengupta, N., & Grober, J. (2005). Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. *Journal of Rheumatology, 32*, 811-819.

Chandran, V., Bhella, S., Schentag, C., & Gladman, D. D. (2007). Functional assessment of chronic illness therapy-fatigue scale is valid in patients with psoriatic arthritis. *Annals of the Rheumatic Diseases, 66*(7), 936-939.

Cleeland, C. S., & Ryan, K. M. (1994). Pain Assessment: Global use of the brief pain inventory. *Annals Academy of Medicine, 23*(2), 129-138.

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research, 16*, 85–94.

Fries, J. F., Spitz, P., Kraines, R. G., & Holman, H. R. (1980). Measurement of patient outcome in arthritis. *Arthritis and Rheumatism, 23*(2), 137-145.

Hagell, P., Hoglund, A., Reimer, J., Eriksson, B., Knutsson, I., Widner, H., & Cella, D. (2006). Measuring fatigue in Parkinson's disease: A psychometric study of two brief generic fatigue questionnaires. *Journal of Pain and Symptom Management, 32*(5), 420-432.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating*. ACT Research Report 94-4. Iowa City, IA: American College Testing.

Kessler, R. C., Barker, P.R., Colpe, L.J., Epstein, J.F., Gfroerer, J.C., Hiripi, E., . . . Zaslavsky, A.M. (2003). Screening for Serious Mental Illness in the General Population. *Archives of General Psychiatry, 60*(2), 184-189.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking : methods and practices*. New York: Springer.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9 Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine, 16*(9), 606-613.

Lord, F. M. (1982). The Standard Error of Equipercentile Equating. *Journal of Educational and Behavioral Statistics, 7*(3), 165-174.

Mease, P. J., Revicki, D. A., Szechinski, J., Greenwald, M., Kivitz, A., Barile-Fabris, L., . . . Leirisalo-Repo, M. (2008). Improved health-related quality of life for patients with active rheumatoid arthritis receiving rituximab: Results of the Dose-Ranging Assessment: International Clinical Evaluation of Rituximab in Rheumatoid Arthritis (DANCER) Trial. *Journal of Rheumatology, 35*(1), 20-30.

Mittendorf, T., Dietz, B., Sterz, R., Kupper, H., Cifaldi, M. A., & von der Schulenburg, J.-M. (2007). Improvement and longterm maintenance of quality of life during treatment with adalimumab in severe rheumatoid arthritis. *Journal of Rheumatology, 34*(12), 2343-2350.

Mulrooney, D. A., Neglia, J. P., Ness, K. K., Robison, L. L., Whitton, J. A., Green, D. M., . . . Mertens, A. C. (2008). Fatigue and sleep disturbance in adult survivors of childhood cancer: A report from the childhood cancer survivor study (CCSS). *Sleep, 31*(2), 271-281.

Ng, A. K., Li, S., Recklitis, C., Neuberg, D., Chakrabarti, S., Silver, B., & Diller, L. (2005). A comparison between long-term survivors of Hodgkin's disease and their siblings on fatigue level and factors predicting for increased fatigue. *Annals of Oncology, 16*(12), 1949-1955.

Quirt, I., Robeson, C., Lau, C. Y., Kovacs, M., Burdette-Radoux, S., Dolan, S., . . . Couture, F. (2001). Epoetin alfa therapy increases hemoglobin levels and improves quality of life in patients with cancer-related anemia who are not receiving chemotherapy and patients with anemia who are receiving chemotherapy. *Journal of Clinical Oncology, 19*(21), 4126-4134.

Quirt, I., Robeson, C., Lau, C. Y., Kovacs, M., Burdette-Radoux, S., Dolan, S., . . . Couture, F. (2002). Epoetin alfa in patients not on chemotherapy - Canadian data. *Seminars in Oncology, 29*(3), 75-80.

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement, 1*(3), 385-401.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik, 10*(3), 177-183.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Chicago, Illinois: Psychometric Society.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine, 166*(10), 1092-1097.

Ware, J. E., Kosinski, M., & Dewey, J. E. (2000). *How to score version 2 of the SF-36 health survey*. Lincoln, R.I.: QualityMetric Inc.

Watson, D., Clark, L. A., Weber, K., Assenheimer, J. S., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology, 104*(1), 15-25.

Weeks, J. P. (2010). `Plink`: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1-33.

Yellen, S. B., Cella, D. F., Webster, K., Blendowski, C., & Kaplan, E. (1997). Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. *Journal of Pain and Symptom Management, 13*(2), 63-74.

7. Appendix

Appendix Table 1: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for MASQ to PROMIS Anxiety (PROMIS Study). RECOMMENDED.

MASQ Score	PROMIS T-score	SE	MASQ Score	PROMIS T-score	SE
11	35.2	6.1	52	86.4	2.6
12	38.9	5.4	53	86.9	2.4
13	41.9	5.0	54	87.4	2.2
14	44.3	4.6	55	87.7	1.9
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			
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			

Appendix Table 2: Direct (Raw to Scale) Equipercentile Crosswalk Table - From MASQ to PROMIS Anxiety (PROMIS Study). Note: Table 1 is recommended.

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
11	32	33	34	0.58
12	39	39	39	0.58
13	43	43	42	0.56
14	45	45	45	0.33
15	47	47	47	0.31
16	49	49	49	0.39
17	51	51	51	0.28
18	52	52	52	0.51
19	54	54	54	0.37
20	55	55	55	0.63
21	56	56	56	0.31
22	57	57	57	0.68
23	58	58	58	0.37
24	60	59	59	0.51
25	60	60	60	0.45
26	61	61	61	0.60
27	62	62	62	0.47
28	62	63	63	0.48
29	63	64	64	0.65
30	64	64	65	0.42
31	65	65	66	3.02
32	66	66	67	0.57
33	67	67	68	1.81
34	68	68	69	0.67
35	69	69	70	1.13
36	71	70	71	1.73
37	72	71	72	1.58
38	73	72	72	0.57
39	73	73	73	0.47
40	73	73	74	0.47
41	73	74	75	0.33
42	77	75	76	0.33
43	83	76	77	1.41
44	84	78	78	0.01
45	84	79	79	0.01
46	84	80	80	0.01
47	84	81	81	0.01
48	84	82	82	0.01
49	84	83	83	0.01
50	85	84	85	0.01
51	86	85	86	0.01
52	87	87	87	0.01

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
53	88	88	88	0.01
54	89	89	89	0.01
55	90	90	90	0.01

Appendix Table 3: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From MASQ to PROMIS Anxiety (PROMIS Study). Note: Table 1 is recommended.

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
11	33	32	31
12	40	39	39
13	43	43	43
14	45	45	46
15	47	47	48
16	49	49	49
17	51	51	51
18	52	52	52
19	54	54	53
20	55	55	54
21	56	56	56
22	57	57	57
23	58	58	58
24	60	59	59
25	60	60	60
26	61	61	61
27	62	62	62
28	62	63	63
29	63	64	64
30	64	64	65
31	65	65	66
32	66	66	66
33	67	67	68
34	68	68	68
35	69	69	69
36	71	70	70
37	72	71	71
38	73	72	72
39	73	73	73
40	73	74	74
41	73	74	74
42	79	74	75
43	83	75	75
44	83	76	76
45	83	77	77
46	83	78	78
47	83	78	79
48	83	79	79
49	84	80	80
50	85	81	82
51	86	83	83
52	86	84	84

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
53	88	85	86
54	88	87	87
55	88	88	88

Appendix Table 4: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for SF-36/MH to PROMIS Anxiety (PROMIS Study). RECOMMENDED.

SF-36/MH Score	PROMIS T-score	SE
25	33.2	6.3
24	38.9	5.6
23	43.1	5.0
22	46.8	4.7
21	49.4	4.5
20	51.7	4.3
19	53.7	4.1
18	55.6	4.0
17	57.3	3.9
16	59.0	3.8
15	60.6	3.8
14	62.1	3.7
13	63.7	3.7
12	65.2	3.7
11	66.8	3.7
10	68.5	3.7
9	70.3	3.8
8	72.2	3.8
7	74.3	3.9
6	76.8	4.0
5	80.1	4.2

Appendix Table 5: Direct (Raw to Scale) Equipercentile Crosswalk Table - From SF-36/MH to PROMIS Anxiety (PROMIS Study). Note: Table 4 is recommended.

SF-36/MH Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
25	32	32	33	0.37
24	38	38	38	0.37
23	43	43	43	0.71
22	48	47	47	0.69
21	51	50	50	0.31
20	53	52	52	0.59
19	54	54	54	0.37
18	56	56	56	0.36
17	57	57	57	0.75
16	58	58	59	0.38
15	60	60	60	0.52
14	62	61	62	0.68
13	63	63	63	0.77
12	64	65	65	0.47
11	66	66	66	0.60
10	68	68	68	0.70
9	69	69	70	0.96
8	71	71	71	1.87
7	73	73	73	0.66
6	79	80	80	1.41
5	88	87	87	2.45

Appendix Table 6: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From SF-36/MH to PROMIS Anxiety (PROMIS Study). Note: Table 4 is recommended.

SF-36/MH Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
25	32	25	10
24	38	39	38
23	43	44	44
22	48	47	47
21	51	50	50
20	52	52	52
19	54	54	54
18	55	55	55
17	57	57	57
16	58	58	58
15	60	60	60
14	62	61	61
13	63	63	63
12	64	64	64
11	66	66	66
10	68	68	68
9	69	69	69
8	71	71	71
7	73	73	73
6	79	78	78
5	88	84	84

Appendix Table 7: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for CES-D to PROMIS Depression (PROMIS Study). RECOMMENDED

CES-D Score	PROMIS T-score	SE	CES-D Score	PROMIS T-score	SE
0	34.5	6.0	40	69.2	2.3
1	38.6	5.1	41	69.8	2.3
2	41.1	4.7	42	70.4	2.3
3	42.9	4.6	43	71.0	2.4
4	44.7	4.1	44	71.7	2.4
5	46.2	3.8	45	72.3	2.5
6	47.5	3.6	46	73.0	2.5
7	48.7	3.4	47	73.7	2.6
8	49.8	3.2	48	74.4	2.7
9	50.8	3.0	49	75.2	2.7
10	51.7	2.9	50	76.0	2.8
11	52.6	2.8	51	76.8	2.9
12	53.4	2.7	52	77.7	3.0
13	54.1	2.6	53	78.7	3.1
14	54.8	2.5	54	79.7	3.2
15	55.5	2.4	55	80.8	3.2
16	56.2	2.4	56	82.0	3.2
17	56.8	2.3	57	83.1	3.2
18	57.4	2.3	58	84.3	3.1
19	58.0	2.3	59	85.4	2.8
20	58.6	2.3	60	86.4	2.5
21	59.1	2.2			
22	59.7	2.2			
23	60.2	2.2			
24	60.8	2.2			
25	61.3	2.2			
26	61.8	2.2			
27	62.3	2.1			
28	62.9	2.1			
29	63.4	2.1			
30	63.9	2.1			
31	64.4	2.1			
32	64.9	2.1			
33	65.4	2.1			
34	66.0	2.2			
35	66.5	2.2			
36	67.0	2.2			
37	67.6	2.2			
38	68.1	2.2			
39	68.7	2.2			

Appendix Table 8: Direct (Raw to Scale) Equipercentile Crosswalk Table - From CES-D to PROMIS Depression (PROMIS Study). Note: Table 7 is recommended.

CES-D Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	34	33	35	0.21
1	38	38	38	0.21
2	40	41	41	0.31
3	44	43	43	0.76
4	46	45	45	0.34
5	47	47	47	0.35
6	48	48	48	0.43
7	50	49	49	0.28
8	50	50	50	0.25
9	51	51	51	0.56
10	52	52	52	0.30
11	52	52	52	0.27
12	53	53	53	0.33
13	54	54	54	0.32
14	54	54	54	0.30
15	55	55	55	0.49
16	56	56	56	0.39
17	56	56	56	0.37
18	57	57	57	0.64
19	58	57	58	0.29
20	58	58	58	0.28
21	58	58	59	0.27
22	59	59	59	0.30
23	59	60	60	0.29
24	60	60	60	0.79
25	61	61	61	0.36
26	61	61	62	0.32
27	62	62	62	0.40
28	62	63	63	0.38
29	63	63	63	0.57
30	64	64	64	0.82
31	64	64	65	0.74
32	65	65	65	1.02
33	66	66	66	0.68
34	66	66	66	0.63
35	67	67	67	1.29
36	68	67	67	0.52
37	68	68	68	0.47
38	68	68	69	0.42
39	69	69	69	1.28
40	69	69	70	1.09
41	70	70	70	1.15

CES-D Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
42	70	71	71	1.15
43	72	71	71	1.54
44	72	72	72	1.46
45	73	73	73	0.79
46	73	73	73	0.80
47	74	74	74	1.84
48	75	74	74	2.45
49	76	75	75	2.00
50	76	76	76	2.00
51	77	77	77	2.00
52	78	79	78	2.00
53	78	80	80	2.00
54	80	81	81	1.41
55	80	83	83	1.41
56	81	84	84	1.41
57	86	86	86	1.41
58	88	87	87	1.41
59	89	88	88	1.41
60	90	90	90	1.41

Appendix Table 9: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From CES-D to PROMIS Depression (PROMIS Study). Note: Table 7 is recommended.

CES-D Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	34	31	28
1	38	38	39
2	41	41	42
3	44	44	44
4	46	46	46
5	47	47	47
6	49	48	48
7	50	50	49
8	50	50	50
9	51	51	51
10	52	52	52
11	52	52	52
12	53	53	53
13	54	54	54
14	54	54	54
15	55	55	55
16	56	56	56
17	56	56	56
18	57	57	57
19	58	58	57
20	58	58	58
21	59	58	58
22	59	59	59
23	59	60	60
24	60	60	60
25	61	61	61
26	61	61	61
27	62	62	62
28	62	62	62
29	63	63	63
30	64	64	64
31	64	64	64
32	65	65	65
33	66	66	66
34	66	66	66
35	67	67	67
36	68	68	68
37	68	68	68
38	68	68	69
39	69	69	69
40	69	69	70
41	69	70	70

CES-D Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
42	70	70	71
43	71	71	72
44	72	72	72
45	73	72	73
46	73	73	73
47	74	74	74
48	75	74	75
49	76	75	76
50	76	76	76
51	77	77	77
52	77	77	77
53	78	78	78
54	80	79	79
55	80	79	79
56	81	80	80
57	86	81	81
58	87	83	83
59	87	84	84
60	87	86	86

Appendix Table 10: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for SF-36/MH to PROMIS Depression (PROMIS Study). RECOMMENDED

SF-36/MH Score	PROMIS T-score	SE
25	33.6	6.0
24	39.6	5.2
23	44.0	4.6
22	47.5	4.3
21	50.1	4.2
20	52.3	3.9
19	54.3	3.7
18	56.1	3.6
17	57.8	3.5
16	59.4	3.4
15	61.0	3.3
14	62.5	3.3
13	64.0	3.3
12	65.5	3.3
11	67.1	3.4
10	68.8	3.4
9	70.6	3.5
8	72.5	3.6
7	74.6	3.8
6	77.1	4.0
5	79.8	4.2

Appendix Table 11: Direct (Raw to Scale) Equipercentile Crosswalk Table - From SF-36/MH to PROMIS Depression (PROMIS Study). Note: Table 10 is recommended.

SF-36/MH Score	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
25	32	32	0.21
24	38	39	0.21
23	44	43	1.05
22	48	47	0.48
21	51	50	0.72
20	53	53	0.41
19	55	55	0.35
18	56	56	0.47
17	58	58	0.36
16	59	59	0.35
15	61	61	0.40
14	62	62	0.50
13	64	64	0.68
12	65	66	1.46
11	67	67	1.57
10	69	69	1.70
9	71	71	1.70
8	73	73	0.71
7	75	75	4.69
6	81	81	2.45
5	87	87	2.45

Appendix Table 12: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From SF-36/MH to PROMIS Depression (PROMIS Study). Note: Table 10 is recommended.

SF-36/MH Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
25	34	34	34
24	39	40	38
23	44	44	44
22	48	48	48
21	51	51	50
20	53	53	52
19	54	54	54
18	56	56	56
17	58	58	58
16	59	59	59
15	61	60	61
14	62	62	62
13	63	63	64
12	65	65	65
11	67	67	67
10	69	69	69
9	71	71	71
8	73	73	73
7	74	75	75
6	78	78	78
5	86	83	82

Appendix Table 13: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for BPAQ to PROMIS Anger (PROMIS Study).

BPAQ Score	PROMIS T-score	SE	BPAQ Score	PROMIS T-score	SE
12	27.3	6.4	52	76.6	4.6
13	28.9	6.5	53	78.0	4.6
14	30.6	6.4	54	79.3	4.5
15	32.0	6.5	55	80.7	4.4
16	33.5	6.3	56	82.0	4.2
17	35.2	6.2	57	83.3	3.9
18	36.7	6.1	58	84.4	3.6
19	38.2	5.9	59	85.3	3.3
20	39.7	5.8	60	86.0	3.0
21	41.1	5.7			
22	42.5	5.5			
23	43.8	5.4			
24	45.1	5.3			
25	46.4	5.2			
26	47.7	5.1			
27	48.9	5.0			
28	50.1	4.9			
29	51.2	4.9			
30	52.4	4.8			
31	53.5	4.7			
32	54.6	4.7			
33	55.7	4.6			
34	56.7	4.6			
35	57.8	4.5			
36	58.8	4.5			
37	59.8	4.5			
38	60.9	4.5			
39	61.9	4.5			
40	62.9	4.5			
41	63.9	4.5			
42	65.0	4.5			
43	66.0	4.5			
44	67.1	4.5			
45	68.2	4.5			
46	69.3	4.5			
47	70.4	4.5			
48	71.6	4.6			
49	72.8	4.6			
50	74.0	4.6			
51	75.3	4.6			

Note. The precision of cross-walk tables is affected by the association between measures. Because the instruments used for these crosswalks were based on measures with correlations less than .80, we recommend caution when using these cross-walk tables.

Appendix Table 14: Direct (Raw to Scale) Equipercentile Crosswalk Table - From BPAQ to PROMIS Anger (PROMIS Study).

BPAQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
12	28	14	14	0.07
13	28	23	24	0.07
14	28	27	28	0.08
15	28	29	30	0.17
16	34	33	32	0.39
17	36	35	35	0.47
18	37	37	37	0.44
19	38	39	39	0.51
20	40	40	40	0.30
21	42	42	42	0.30
22	43	43	43	0.74
23	44	44	44	0.37
24	45	45	45	0.86
25	46	46	46	0.42
26	48	48	48	0.50
27	49	49	49	0.50
28	50	50	50	0.26
29	50	51	51	0.24
30	52	52	52	0.46
31	53	53	53	0.57
32	54	54	54	0.44
33	54	54	54	0.38
34	55	55	55	0.46
35	56	56	56	0.52
36	57	57	57	0.55
37	58	58	58	0.34
38	58	58	59	0.32
39	59	59	60	0.80
40	60	60	61	0.54
41	61	61	62	1.13
42	62	62	63	1.18
43	64	63	64	0.72
44	64	64	65	0.64
45	65	65	66	0.81
46	66	67	67	1.89
47	68	68	68	1.63
48	69	69	69	0.85
49	70	71	70	0.98
50	74	72	72	0.77
51	74	74	74	2.00
52	74	76	76	2.00

BPAQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
53	74	78	77	2.00
54	75	79	79	2.00
55	79	81	81	1.41
56	86	83	83	1.41
57	87	84	84	1.41
58	88	86	86	1.41
59	89	88	88	1.41
60	90	90	90	1.41

Note. The precision of cross-walk tables is affected by the association between measures. Because the instruments used for these crosswalks were based on measures with correlations less than .80, we recommend caution when using these cross-walk tables.

Appendix Table 15: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From BPAQ to PROMIS Anger (PROMIS Study)

BPAQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
12	26	23	20
13	26	21	14
14	28	26	23
15	29	30	29
16	33	33	33
17	35	35	35
18	37	38	38
19	39	39	39
20	40	40	41
21	42	42	42
22	43	43	43
23	44	44	44
24	45	46	46
25	46	47	47
26	48	48	48
27	49	49	49
28	50	50	50
29	51	51	51
30	51	52	52
31	53	53	52
32	54	54	53
33	54	54	54
34	55	55	55
35	56	56	56
36	57	57	57
37	58	58	58
38	58	59	59
39	59	59	60
40	60	60	60
41	61	61	61
42	62	62	62
43	64	63	63
44	64	64	64
45	65	65	65
46	66	66	66
47	68	68	68
48	69	69	68
49	70	70	70
50	74	72	71
51	74	73	73
52	74	74	74
53	74	76	76

BPAQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
54	75	78	77
55	79	79	79
56	88	81	81
57	88	83	83
58	88	85	85
59	89	87	87
60	89	89	89

Note. The precision of cross-walk tables is affected by the association between measures. Because the instruments used for these crosswalks were based on measures with correlations less than .80, we recommend caution when using these cross-walk tables.

Appendix Table 16: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for HAQ-DI to PROMIS Physical Function (PROMIS Study) RECOMMENDED

Estimated HAQ-DI Score*	PROsetta HAQ-DI Score	PROMIS T-score	SE	Estimated HAQ-DI Score *	PROsetta HAQ-DI Score	PROMIS T-score	SE
2.65	20	12.5	1.7	0.80	57	33.9	1.5
2.60	21	13.4	2.0	0.75	58	34.4	1.5
2.55	22	14.2	2.1	0.70	59	35.0	1.6
2.50	23	15.1	2.2	0.65	60	35.5	1.6
2.45	24	16.0	2.1	0.60	61	36.1	1.6
2.40	25	16.9	2.1	0.55	62	36.7	1.6
2.35	26	17.7	2.0	0.50	63	37.4	1.7
2.30	27	18.4	1.9	0.45	64	38.1	1.7
2.25	28	19.1	1.8	0.40	65	38.8	1.8
2.20	29	19.8	1.8	0.35	66	39.6	1.8
2.15	30	20.4	1.7	0.30	67	40.4	1.9
2.10	31	21.0	1.7	0.25	68	41.4	2.0
2.05	32	21.6	1.6	0.20	69	42.5	2.2
2.00	33	22.1	1.6	0.15	70	43.9	2.6
1.95	34	22.7	1.6	0.10	71	45.7	2.9
1.90	35	23.2	1.6	0.05	72	48.6	3.8
1.85	36	23.7	1.5	0.00	73	56.8	6.8
1.80	37	24.2	1.5				
1.75	38	24.7	1.5				
1.70	39	25.2	1.5				
1.65	40	25.7	1.5				
1.60	41	26.1	1.5				
1.55	42	26.6	1.5				
1.50	43	27.1	1.5				
1.45	44	27.5	1.5				
1.40	45	28.0	1.5				
1.35	46	28.5	1.5				
1.30	47	28.9	1.5				
1.25	48	29.4	1.5				
1.20	49	29.9	1.5				
1.15	50	30.4	1.5				
1.10	51	30.8	1.5				
1.05	52	31.3	1.5				
1.00	53	31.8	1.5				
0.95	54	32.3	1.5				
0.90	55	32.8	1.5				
0.85	56	33.3	1.5				

*The HAQ-DI scores were estimated from PROSetta raw summed scores as follows: ((73-x)/53)*2.65. Higher HAQ-DI scores indicate more disability.

Appendix Table 17: Direct (Raw to Scale) Equipercentile Crosswalk Table – From HAQ to PROMIS Physical Function (PROMIS Study). Note: Table 16 is recommended.

Estimated HAQ-DI Score *	PROsetta HAQ-DI Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
2.65	20	10	10	10	1.41
2.60	21	11	11	11	1.41
2.55	22	12	11	11	1.41
2.50	23	13	12	12	1.41
2.45	24	14	13	13	1.41
2.40	25	15	14	14	1.41
2.35	26	16	15	15	1.41
2.30	27	17	15	15	1.41
2.25	28	18	16	16	1.41
2.20	29	18	17	17	1.41
2.15	30	18	18	18	1.41
2.10	31	18	18	18	1.41
2.05	32	18	19	19	1.41
2.00	33	22	20	20	1.41
1.95	34	22	21	21	1.41
1.90	35	22	22	21	1.41
1.85	36	22	22	22	1.41
1.80	37	24	23	23	0.61
1.75	38	25	24	24	0.35
1.70	39	25	24	24	0.35
1.65	40	25	25	25	0.35
1.60	41	25	25	25	0.35
1.55	42	25	25	25	0.35
1.50	43	25	25	26	0.35
1.45	44	25	26	26	0.35
1.40	45	28	27	27	0.71
1.35	46	28	28	28	0.71
1.30	47	29	29	28	0.67
1.25	48	30	30	29	1.41
1.20	49	31	31	30	0.72
1.15	50	31	31	30	0.67
1.10	51	31	31	31	0.67
1.05	52	32	32	31	0.41
1.00	53	32	32	32	0.47
0.95	54	32	33	32	0.47
0.90	55	33	33	33	0.54
0.85	56	33	33	33	0.49
0.80	57	33	34	34	0.52
0.75	58	34	34	34	0.87

Estimated HAQ-DI Score *	PROsetta HAQ-DI Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0.70	59	35	35	35	0.81
0.65	60	36	35	35	0.80
0.60	61	36	36	36	0.57
0.55	63	37	36	37	0.44
0.50	63	37	37	37	0.39
0.45	64	37	38	38	0.37
0.40	65	38	38	39	0.49
0.35	66	40	39	40	0.33
0.30	67	40	41	41	0.29
0.25	68	42	42	42	0.27
0.20	69	43	43	43	0.41
0.15	70	44	44	44	0.32
0.10	71	45	45	45	0.39
0.05	72	47	47	47	0.33
0.00	73	57	53	52	2.46

*The HAQ-DI scores were estimated from PROSetta raw summed scores as follows: ((73-x)/53)*2.65. Higher HAQ-DI scores indicate more disability.

Appendix Table 18: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From HAQ to PROMIS Physical Function (PROMIS Study). Note: Table 16 is recommended.

Estimated HAQ-DI Score *	PROsetta HAQ-DI Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
2.65	20	11	11	11
2.60	21	11	12	12
2.55	22	11	13	12
2.50	23	11	14	14
2.45	24	11	15	15
2.40	25	12	16	16
2.35	26	12	17	17
2.30	27	12	18	18
2.25	28	12	19	18
2.20	29	13	20	19
2.15	30	13	20	20
2.10	31	14	21	20
2.05	32	18	21	21
2.00	33	18	22	22
1.95	34	18	22	22
1.90	35	22	23	23
1.85	36	22	24	23
1.80	37	23	24	24
1.75	38	25	24	24
1.70	39	25	25	24
1.65	40	25	25	24
1.60	41	25	25	25
1.55	42	25	25	25
1.50	43	25	25	25
1.45	44	25	25	25
1.40	45	25	26	26
1.35	46	28	28	28
1.30	47	29	29	29
1.25	48	30	30	30
1.20	49	31	31	31
1.15	50	31	31	31
1.10	51	31	32	32
1.05	52	32	32	32
1.00	53	32	32	32
0.95	54	32	32	32
0.90	55	33	33	33
0.85	56	33	33	33
0.80	57	33	34	34
0.75	58	34	34	34
0.70	59	35	35	35

Estimated HAQ-DI Score *	PROsetta HAQ-DI Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0.65	60	35	35	35
0.60	61	36	36	36
0.55	62	37	36	36
0.50	63	37	37	37
0.45	64	37	37	37
0.40	65	38	38	38
0.35	66	40	40	39
0.30	67	40	40	40
0.25	68	42	42	42
0.20	69	43	43	43
0.15	70	44	44	44
0.10	71	45	45	45
0.05	72	47	47	47
0.00	73	56	56	55

*The HAQ-DI scores were estimated from PROSetta raw summed scores as follows: ((73-x)/53)*2.65. Higher HAQ-DI scores indicate more disability.

Appendix Table 19: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for SF-36/PF to PROMIS Physical Function (PROMIS Study). RECOMMENDED.

SF-36/PF Score	PROMIS T-score	SE
10	24.5	4.0
11	28.3	2.8
12	30.3	2.5
13	32.0	2.2
14	33.4	2.1
15	34.8	2.0
16	36.0	2.0
17	37.2	2.0
18	38.4	1.9
19	39.5	1.9
20	40.7	1.9
21	41.8	1.9
22	42.9	1.9
23	44.1	2.0
24	45.3	2.0
25	46.7	2.1
26	48.2	2.3
27	49.9	2.5
28	52.0	2.9
29	55.0	3.5
30	61.7	5.7

Appendix Table 20: Direct (Raw to Scale) Equipercentile Crosswalk Table – From SF-36/PF to PROMIS Physical Function (PROMIS Study). Note: Table 19 is recommended.

SF-36/PF Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
10	24	15	15	3.46
11	28	27	27	1.27
12	29	29	29	0.71
13	31	31	31	1.05
14	33	33	32	0.70
15	34	34	34	1.00
16	36	36	36	0.55
17	37	37	37	0.47
18	39	39	39	0.79
19	41	40	40	0.65
20	42	41	41	0.29
21	42	42	42	0.28
22	43	43	43	0.41
23	44	44	44	0.25
24	46	46	46	0.34
25	47	47	47	0.25
26	48	48	48	0.25
27	50	49	50	0.24
28	51	51	52	0.27
29	55	55	55	0.32
30	62	61	60	1.01

Appendix Table 21: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From SF-36/PF to PROMIS Physical Function (PROMIS Study). Note: Table 19 is recommended.

SF-36/PF Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
10	23	19	19
11	28	28	28
12	29	30	29
13	31	31	31
14	33	33	33
15	34	34	34
16	36	36	36
17	37	37	37
18	39	39	39
19	41	40	40
20	42	42	41
21	42	42	42
22	43	43	43
23	44	44	44
24	45	46	45
25	47	47	47
26	48	48	48
27	50	50	50
28	51	52	52
29	55	55	55
30	62	62	62

Appendix Table 22: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for FACIT-F to PROMIS Fatigue (PROMIS Study) RECOMMENDED

FACIT-F Score	PROMIS T-score	SE	FACIT-F Score	PROMIS T-score	SE
52	30.3	4.8	12	68.9	2.0
51	35.0	3.5	11	69.6	2.0
50	38.0	3.0	10	70.4	2.0
49	40.3	2.8	9	71.2	2.1
48	42.1	2.6	8	72.0	2.2
47	43.7	2.5	7	72.9	2.3
46	45.0	2.3	6	73.9	2.4
45	46.3	2.2	5	75.0	2.5
44	47.3	2.1	4	76.2	2.7
43	48.3	2.0	3	77.5	2.9
42	49.3	2.0	2	79.1	3.1
41	50.1	1.9	1	81.2	3.3
40	51.0	1.9	0	83.5	3.4
39	51.7	1.9			
38	52.5	1.9			
37	53.2	1.9			
36	53.9	1.8			
35	54.6	1.8			
34	55.3	1.8			
33	55.9	1.8			
32	56.6	1.8			
31	57.2	1.8			
30	57.8	1.8			
29	58.4	1.8			
28	59.0	1.8			
27	59.6	1.8			
26	60.2	1.8			
25	60.8	1.8			
24	61.4	1.8			
23	62.0	1.8			
22	62.6	1.8			
21	63.2	1.8			
20	63.8	1.8			
19	64.4	1.8			
18	65.0	1.8			
17	65.6	1.8			
16	66.2	1.9			
15	66.9	1.9			
14	67.5	1.9			
13	68.2	1.9			

Appendix Table 23: Direct (Raw to Scale) Equipercentile Crosswalk Table – From FACIT-F to PROMIS Fatigue (PROMIS Study). Note: Table 22 is recommended.

FACIT-F Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
52	30	32	33	1.28
51	36	36	36	0.60
50	38	38	38	0.32
49	40	40	40	0.27
48	42	42	42	0.21
47	44	44	43	0.24
46	45	45	45	0.53
45	46	46	46	0.19
44	48	47	47	0.24
43	48	49	48	0.21
42	50	49	49	0.33
41	50	50	50	0.24
40	51	51	51	0.26
39	52	52	52	0.29
38	52	53	53	0.26
37	53	53	53	0.28
36	54	54	54	0.30
35	55	55	55	0.41
34	56	56	55	0.33
33	56	56	56	0.26
32	57	57	57	0.21
31	57	57	57	0.20
30	58	58	58	0.18
29	58	58	59	0.20
28	59	59	59	0.30
27	59	60	60	0.26
26	60	60	60	0.29
25	60	61	61	0.45
24	61	61	62	0.39
23	62	62	62	0.39
22	63	63	63	0.70
21	64	63	63	0.33
20	64	64	64	0.30
19	65	64	64	0.56
18	65	65	65	0.45
17	66	65	66	0.36
16	66	66	66	0.32
15	66	66	67	0.30
14	66	67	67	0.31
13	67	67	68	0.53

FACIT-F Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
12	68	68	68	0.35
11	69	69	69	1.09
10	70	70	70	1.09
9	70	70	70	0.75
8	72	71	71	0.82
7	72	72	72	0.67
6	73	74	74	0.67
5	74	77	77	0.47
4	74	79	79	1.41
3	84	82	82	1.41
2	84	84	84	1.41
1	84	87	87	1.41
0	90	89	89	1.41

Appendix Table 24: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From FACIT-F to PROMIS Fatigue (PROMIS Study). Note: Table 22 is recommended.

FACIT-F Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
52	30	30	29
51	36	35	35
50	38	38	38
49	40	40	40
48	42	42	42
47	44	44	44
46	45	45	45
45	46	46	46
44	48	48	47
43	48	48	48
42	50	50	49
41	50	50	50
40	51	51	51
39	52	52	52
38	52	52	53
37	53	53	54
36	54	54	54
35	55	55	55
34	56	56	56
33	56	56	56
32	57	57	57
31	57	57	57
30	58	58	58
29	58	58	58
28	59	59	59
27	59	59	60
26	60	60	60
25	61	61	61
24	61	61	61
23	62	62	62
22	63	63	63
21	64	63	63
20	64	64	64
19	65	64	64
18	65	65	65
17	66	66	66
16	66	66	66
15	66	66	66

FACIT-F Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
14	66	67	67
13	67	67	68
12	68	68	68
11	69	69	69
10	70	70	70
9	70	71	71
8	72	72	72
7	73	73	72
6	73	74	73
5	74	75	74
4	74	76	76
3	84	77	77
2	84	79	79
1	84	82	82
0	88	86	86

Appendix Table 25: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for SF-36/VT to PROMIS Fatigue (PROMIS Study). RECOMMENDED

SF-36/VT Score	PROMIS T-score	SE
20	28.9	4.9
19	34.1	4.1
18	38.1	3.8
17	41.6	3.7
16	44.8	3.5
15	47.5	3.5
14	49.9	3.4
13	52.0	3.4
12	54.1	3.4
11	56.0	3.3
10	57.9	3.3
9	59.9	3.3
8	61.9	3.4
7	64.0	3.4
6	66.4	3.5
5	69.4	3.7
4	74.0	4.5

Appendix Table 26: Direct (Raw to Scale) Equipercentile Crosswalk Table - From SF-36/VT to PROMIS Fatigue (PROMIS Study). Note: Table 25 is recommended.

SF-36/VT Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
20	28	29	29	1.10
19	34	34	34	0.87
18	38	38	38	0.47
17	42	41	41	0.29
16	45	45	45	0.71
15	48	48	48	0.28
14	50	50	50	0.53
13	52	52	52	0.43
12	54	54	54	0.44
11	56	56	56	0.39
10	58	58	58	0.40
9	60	60	60	0.53
8	62	62	62	0.50
7	64	64	64	0.53
6	66	66	66	0.76
5	68	68	68	0.52
4	73	71	71	1.70

Appendix Table 27: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From SF-36/VT to PROMIS Fatigue (PROMIS Study). Note: Table 25 is recommended.

SF-36/VT Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
20	28	26	25
19	34	34	33
18	38	38	38
17	42	42	42
16	45	45	45
15	48	48	48
14	50	50	50
13	52	52	52
12	54	54	54
11	56	56	56
10	58	58	58
9	60	60	60
8	62	62	62
7	64	64	64
6	66	66	66
5	68	69	69
4	73	72	72

Appendix Table 28: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for BPI Severity to PROMIS Pain Interference (PROMIS Study). RECOMMENDED

Actual BPI Severity Score‡	Raw Summed BPI Severity Score	PROMIS T-score	SE
0	4	34.6	5.6
1	5	41.1	4.4
2	6	45.8	4.2
3	7	49.4	4.1
4	8	52.5	4.1
5	9	55.2	4.0
6	10	57.9	4.0
7	11	60.7	4.0
8	12	63.2	4.1
9	13	65.9	4.3
10	14	68.8	4.7
11	15	71.0	4.7
12	16	75.0	5.1

‡ = BPI scores were collapsed from 0-10 to 0-3 where 0 ϵ {0}, 1-4 ϵ {1}, 5-6 ϵ {2}, 7-10 ϵ {3}

Appendix Table 29: Direct (Raw to Scale) Equipercentile Crosswalk Table - From BPI Severity to PROMIS Pain Interference (PROMIS Study). Note: Table 28 is recommended.

Actual BPI Severity Score‡	Raw Summed BPI Severity Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	4	37	31	30	0.04
1	5	37	37	38	0.04
2	6	45	45	45	0.44
3	7	50	50	50	0.42
4	8	53	53	53	0.58
5	9	56	56	56	0.51
6	10	58	58	58	0.47
7	11	60	60	60	0.70
8	12	63	63	63	1.17
9	13	66	65	65	0.55
10	14	67	67	67	0.76
11	15	68	69	69	0.77
12	16	74	75	73	1.25

ǂ = BPI scores were collapsed from 0-10 to 0-3 where 0 ϵ {0}, 1-4 ϵ {1}, 5-6 ϵ {2}, 7-10 ϵ {3}

Appendix Table 30: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From BPI Severity to PROMIS Pain Interference (PROMIS Study). Note: Table 28 is recommended.

Actual BPI Severity Score‡	Raw Summed BPI Severity Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	4	37	37	37
1	5	40	40	37
2	6	45	45	46
3	7	50	50	50
4	8	53	53	53
5	9	56	56	55
6	10	58	58	58
7	11	60	60	60
8	12	63	63	62
9	13	66	65	65
10	14	67	67	67
11	15	68	69	69
12	16	74	74	74

ǂ = BPI scores were collapsed from 0-10 to 0-3 where 0 ϵ {0}, 1-4 ϵ {1}, 5-6 ϵ {2}, 7-10 ϵ {3}

Appendix Table 31: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for BPI Interference to PROMIS Pain Interference (PROMIS Study). RECOMMENDED

Actual BPI Interference Score‡	Raw Summed BPI Interference Score	PROMIS T-score	SE
0	7	38.5	5.7
1	8	44.7	3.5
2	9	47.5	2.7
3	10	49.3	2.6
4	11	51.0	2.4
5	12	52.5	2.3
6	13	53.9	2.3
7	14	55.1	2.3
8	15	56.2	2.3
9	16	57.3	2.3
10	17	58.3	2.2
11	18	59.3	2.1
12	19	60.2	2.1
13	20	61.2	2.0
14	21	62.1	2.0
15	22	63.0	2.0
16	23	63.9	2.1
17	24	65.0	2.2
18	25	66.1	2.4
19	26	67.6	2.8
20	27	68.8	2.8
21	28	72.7	4.2

ǂ = BPI scores were collapsed from 0-10 to 0-3 where 0 ϵ {0}, 1-4 ϵ {1}, 5-6 ϵ {2}, 7-10 ϵ {3}

Appendix Table 32: Direct (Raw to Scale) Equipercentile Crosswalk Table - From BPI Interference to PROMIS Pain Interference (PROMIS Study). Note: Table 31 is recommended.

Actual BPI Interference Score‡	Raw Summed BPI Interference Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	7	37	37	37	0.33
1	8	45	45	45	0.33
2	9	49	48	48	0.29
3	10	50	50	50	0.26
4	11	51	51	51	0.26
5	12	52	52	52	0.21
6	13	53	53	54	0.30
7	14	55	55	55	0.33
8	15	57	56	56	0.39
9	16	57	57	57	0.30
10	17	58	58	58	0.27
11	18	59	59	59	0.60
12	19	60	60	60	0.39
13	20	62	61	61	0.34
14	21	62	62	62	0.31
15	22	62	63	63	0.29
16	23	63	63	63	0.66
17	24	64	64	64	0.55
18	25	66	65	65	0.40
19	26	66	66	67	0.36
20	27	67	68	68	0.51
21	28	70	69	70	3.87

‡ = BPI scores were collapsed from 0-10 to 0-3 where 0 ϵ {0}, 1-4 ϵ {1}, 5-6 ϵ {2}, 7-10 ϵ {3}

Appendix Table 33: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From BPI Interference to PROMIS Pain Interference (PROMIS Study). Note: Table 31 is recommended.

Actual BPI Interference Score‡	Raw Summed BPI Interference Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	7	38	37	33
1	8	45	46	46
2	9	49	48	48
3	10	50	50	50
4	11	51	51	51
5	12	52	52	53
6	13	53	54	54
7	14	55	55	55
8	15	56	56	56
9	16	57	57	57
10	17	58	58	58
11	18	59	59	59
12	19	60	60	60
13	20	62	61	61
14	21	62	62	62
15	22	62	63	63
16	23	63	64	64
17	24	64	64	64
18	25	66	65	65
19	26	66	66	66
20	27	67	67	67
21	28	70	69	69

‡ = BPI scores were collapsed from 0-10 to 0-3 where 0 ϵ {0}, 1-4 ϵ {1}, 5-6 ϵ {2}, 7-10 ϵ {3}

Appendix Table 34: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for GAD-7 to PROMIS Anxiety (Toolbox Study). RECOMMENDED

GAD-7 Score	PROMIS 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
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

Appendix Table 35: Direct (Raw to Scale) Equipercentile Crosswalk Table - From GAD-7 to PROMIS Anxiety (Toolbox Study). Note: Table 34 is recommended.

GAD-7 Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	39	39	40	0.74
1	47	46	46	0.28
2	49	49	49	0.29
3	50	51	51	0.67
4	53	53	53	0.31
5	55	55	55	0.67
6	57	57	57	0.63
7	59	59	58	0.61
8	60	60	60	0.54
9	62	61	61	0.33
10	62	62	62	0.27
11	63	63	63	0.57
12	64	64	64	0.46
13	65	65	66	0.55
14	66	67	67	0.68
15	68	68	68	1.13
16	69	69	70	1.19
17	71	71	72	0.68
18	73	73	74	1.23
19	76	76	76	1.87
20	78	79	78	0.63
21	82	87	86	1.27

Appendix Table 36: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From GAD-7 to PROMIS Anxiety (Toolbox Study). Note: Table 34 is recommended.

GAD-7 Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	38	39	40
1	46	46	46
2	48	49	49
3	50	51	51
4	53	53	53
5	54	55	55
6	57	57	57
7	59	59	58
8	60	60	60
9	62	61	61
10	62	62	62
11	63	63	63
12	64	64	64
13	65	65	66
14	66	66	67
15	68	68	68
16	69	69	70
17	71	71	71
18	73	73	73
19	76	75	75
20	78	78	77
21	83	84	83

Appendix Table 37: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for K6 to PROMIS Anxiety (Toolbox Study)

K6 Score	PROMIS T-score	SE
30	36.2	6.6
29	41.1	5.8
28	44.3	5.5
27	47.0	5.3
26	49.2	5.3
25	51.2	5.1
24	53.1	5.0
23	54.8	4.9
22	56.5	4.8
21	58.0	4.7
20	59.5	4.6
19	61.0	4.6
18	62.4	4.5
17	63.8	4.5
16	65.2	4.5
15	66.6	4.5
14	67.9	4.5
13	69.4	4.5
12	70.8	4.5
11	72.3	4.5
10	73.9	4.6
9	75.6	4.6
8	77.4	4.6
7	79.3	4.5
6	81.8	4.4

Note. The precision of cross-walk tables is affected by the association between measures. Because the instruments used for these crosswalks were based on measures with correlations less than .80, we recommend caution when using these cross-walk tables.

Appendix Table 38: Direct (Raw to Scale) Equipercentile Crosswalk Table - From K6 to PROMIS Anxiety (Toolbox Study)

K6 Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
30	32	33	34	0.53
29	40	40	40	0.53
28	45	44	44	1.02
27	47	48	48	0.26
26	50	50	50	0.81
25	53	53	52	0.33
24	54	54	54	0.51
23	56	56	56	0.56
22	57	57	57	0.60
21	58	58	58	0.51
20	59	59	59	0.58
19	61	61	61	0.47
18	62	62	62	0.44
17	63	63	63	0.70
16	64	64	65	0.62
15	65	65	66	0.68
14	66	67	67	0.85
13	69	69	69	1.41
12	71	70	71	0.75
11	72	72	72	1.47
10	73	74	74	1.31
9	76	76	76	2.45
8	78	77	77	0.92
7	78	79	79	0.67
6	82	81	81	2.35

Note. The precision of cross-walk tables is affected by the association between measures. Because the instruments used for these crosswalks were based on measures with correlations less than .80, we recommend caution when using these cross-walk tables.

Appendix Table 39: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From K6 to PROMIS Anxiety (Toolbox Study)

K6 Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
30	34	33	35
29	41	41	42
28	45	45	45
27	48	48	48
26	50	50	50
25	53	52	52
24	54	54	53
23	55	55	55
22	57	57	56
21	58	58	58
20	59	59	59
19	60	60	60
18	62	62	62
17	63	63	63
16	64	64	65
15	65	66	66
14	66	67	68
13	69	68	69
12	71	70	70
11	72	72	72
10	73	74	74
9	75	75	75
8	77	77	77
7	78	79	79
6	82	82	81

Note. The precision of cross-walk tables is affected by the association between measures. Because the instruments used for these crosswalks were based on measures with correlations less than .80, we recommend caution when using these cross-walk tables.

Appendix Table 40: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for MASQ to PROMIS Anxiety (Toolbox Study). RECOMMENDED

MASQ Score	PROMIS T-score	SE	MASQ Score*	PROMIS T-score	SE
28	33.8	5.9	68	67.0	2.2
29	37.6	5.1	69	67.4	2.2
30	40.1	4.7	70	67.8	2.2
31	42.0	4.5	71	68.2	2.2
32	43.6	4.2	72	68.6	2.2
33	45.1	3.9	73	69.0	2.2
34	46.5	3.7	74	69.4	2.2
35	47.7	3.5	75	69.8	2.2
36	48.8	3.4	76	70.2	2.2
37	49.8	3.2	77	70.6	2.2
38	50.7	3.1	78	70.9	2.2
39	51.6	3.0	79	71.3	2.2
40	52.4	2.9	80	71.7	2.2
41	53.1	2.9	81	72.1	2.2
42	53.9	2.8	82	72.4	2.2
43	54.6	2.7	83	72.8	2.2
44	55.2	2.7	84	73.2	2.2
45	55.9	2.7	85	73.6	2.2
46	56.5	2.6	86	74.0	2.2
47	57.1	2.6	87	74.3	2.2
48	57.6	2.6	88	74.7	2.2
49	58.2	2.5	89	75.1	2.2
50	58.8	2.5	90	75.5	2.2
51	59.3	2.5	91	75.8	2.2
52	59.8	2.5	92	76.2	2.2
53	60.3	2.4	93	76.6	2.2
54	60.8	2.4	94	77.0	2.2
55	61.3	2.4	95	77.4	2.2
56	61.8	2.4	96	77.8	2.3
57	62.3	2.4	97	78.1	2.3
58	62.7	2.3	98	78.5	2.3
59	63.2	2.3	99	78.9	2.3
60	63.7	2.3	100	79.3	2.3
61	64.1	2.3	101	79.7	2.3
62	64.5	2.3	102	80.1	2.3
63	65.0	2.3	103	80.5	2.3
64	65.4	2.3	104	80.9	2.3
65	65.8	2.3	105	81.4	2.3
66	66.2	2.2	106	81.8	2.4
67	66.6	2.2	107	82.2	2.4

MASQ Score	PROMIS T-score	SE
108	82.6	2.4
109	83.0	2.4
110	83.5	2.4
111	83.9	2.4
112	84.3	2.4
113	84.7	2.3
114	85.1	2.3
115	85.5	2.3
116	85.9	2.2
117	86.2	2.1
118	86.6	2.1
119	86.9	2.0
120	87.2	1.9
121	87.4	1.8
122	87.7	1.7
123	87.9	1.6
124	88.1	1.5
125	88.2	1.4
126	88.4	1.3
127	88.5	1.3
128	88.6	1.2
129	88.7	1.1
130	88.8	1.0
131	88.9	1.0
132	89.0	0.9
133	89.1	0.9
134	89.1	0.8
135	89.2	0.8
136	89.3	0.7
137	89.3	0.7
138	89.3	0.7
139	89.4	0.6
140	89.4	0.6

Appendix Table 41: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From MASQ to PROMIS Anxiety (Toolbox Study). Note: Table 40 is recommended.

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
28	32	33	32	0.36
29	36	37	37	0.36
30	40	40	40	0.49
31	42	42	42	0.30
32	44	44	44	1.12
33	46	46	46	0.82
34	47	47	47	0.25
35	49	48	48	0.31
36	49	50	50	0.29
37	50	51	51	0.64
38	52	52	52	0.58
39	53	53	53	0.33
40	54	54	54	0.50
41	54	55	54	0.47
42	55	55	55	0.61
43	56	56	56	0.49
44	56	57	56	0.48
45	57	57	57	0.55
46	58	58	57	0.50
47	58	58	58	0.48
48	59	59	58	0.53
49	59	59	59	0.51
50	60	60	59	0.57
51	60	60	60	0.56
52	60	60	60	0.54
53	61	61	61	0.37
54	61	61	61	0.35
55	61	61	61	0.35
56	61	62	62	0.35
57	62	62	62	0.32
58	62	62	62	0.30
59	62	62	63	0.29
60	62	63	63	0.29
61	63	63	63	0.58
62	63	63	64	0.58
63	64	64	64	0.59
64	64	64	64	0.54
65	64	64	65	0.52
66	65	65	65	0.60
67	65	65	65	0.59
68	65	66	66	0.59
69	66	66	66	0.74

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
70	67	67	67	1.07
71	67	67	67	0.99
72	68	67	68	1.07
73	68	68	68	1.07
74	69	68	68	1.02
75	69	69	69	1.02
76	70	69	69	1.70
77	70	70	70	1.76
78	70	70	70	1.76
79	70	70	70	1.76
80	71	71	71	0.66
81	71	71	71	0.66
82	71	71	71	0.64
83	71	72	72	0.62
84	72	72	72	1.34
85	72	73	73	1.19
86	73	73	73	1.17
87	73	74	74	1.08
88	74	74	74	4.24
89	75	75	74	1.97
90	76	75	75	1.97
91	76	76	75	1.97
92	76	76	76	1.97
93	77	76	76	3.74
94	78	77	76	4.00
95	78	77	77	0.80
96	78	78	77	0.70
97	78	78	78	0.67
98	78	78	78	0.66
99	78	79	78	0.63
100	78	79	79	0.63
101	80	79	79	1.05
102	80	79	79	0.86
103	80	80	80	0.86
104	80	80	80	0.86
105	80	80	80	0.86
106	80	80	80	0.86
107	80	81	81	0.86
108	80	81	81	0.86
109	80	81	81	0.86
110	81	81	82	2.45
111	82	82	82	2.45
112	82	82	82	1.22
113	82	82	82	1.00

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
114	82	82	82	1.00
115	82	83	83	1.00
116	82	83	83	1.00
117	82	83	83	1.00
118	82	84	84	1.00
119	83	84	84	2.00
120	84	84	84	1.41
121	84	84	85	1.41
122	84	85	85	1.41
123	84	85	85	1.41
124	84	85	85	1.41
125	87	86	86	0.71
126	87	86	86	0.35
127	87	86	86	0.35
128	87	87	87	0.35
129	87	87	87	0.35
130	87	87	87	0.35
131	87	88	88	0.35
132	87	88	88	0.35
133	87	88	88	0.35
134	87	88	89	0.35
135	87	89	89	0.35
136	87	89	89	0.35
137	87	89	89	0.35
138	87	90	90	0.35
139	87	90	90	0.35
140	90	90	90	0.35

Appendix Table 42: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From MASQ to PROMIS Anxiety (Toolbox Study). Note: Table 40 is recommended.

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
28	33	32	32
29	38	37	37
30	41	40	41
31	42	43	43
32	44	44	45
33	46	46	46
34	47	47	48
35	48	48	49
36	49	50	50
37	51	51	51
38	52	52	52
39	53	53	52
40	54	54	53
41	54	54	54
42	55	55	55
43	56	56	55
44	57	56	56
45	57	57	57
46	58	58	57
47	58	58	58
48	59	59	58
49	59	59	59
50	60	60	59
51	60	60	60
52	60	60	60
53	61	61	60
54	61	61	61
55	61	61	61
56	61	62	62
57	62	62	62
58	62	62	62
59	62	62	63
60	63	63	63
61	63	63	63
62	63	63	64
63	64	64	64
64	64	64	64
65	64	64	65
66	65	65	65
67	65	65	66
68	65	66	66
69	66	66	66

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
70	67	66	67
71	67	67	67
72	68	67	67
73	68	68	68
74	69	68	68
75	69	68	68
76	70	69	69
77	70	69	69
78	70	70	70
79	70	70	70
80	70	71	70
81	71	71	71
82	71	71	71
83	71	72	72
84	72	72	72
85	72	73	72
86	73	73	73
87	73	74	73
88	74	74	74
89	75	74	74
90	75	75	74
91	76	75	75
92	76	76	75
93	77	76	75
94	77	76	76
95	77	77	76
96	78	77	77
97	78	78	77
98	78	78	77
99	78	78	78
100	78	78	78
101	80	79	79
102	80	79	79
103	80	79	80
104	80	80	80
105	80	80	80
106	80	80	81
107	80	81	82
108	80	81	82
109	80	82	82
110	81	82	82
111	82	82	83
112	82	82	83
113	83	82	83

MASQ Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
114	83	82	83
115	83	83	83
116	83	83	84
117	83	83	84
118	83	83	84
119	83	83	84
120	84	84	84
121	84	84	84
122	84	84	84
123	84	84	85
124	84	84	85
125	86	84	85
126	87	85	85
127	87	85	85
128	87	85	85
129	87	85	86
130	87	85	86
131	87	86	86
132	87	86	86
133	87	86	86
134	87	86	86
135	87	86	86
136	87	86	86
137	87	87	87
138	87	87	87
139	87	87	87
140	87	87	87

Appendix Table 43: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for CES-D to PROMIS Depression (Toolbox Study). RECOMMENDED

CES-D Score	PROMIS T-score	SE	CES-D Score	PROMIS T-score	SE
0	32.4	6.0	40	70.4	2.6
1	35.9	5.4	41	71.1	2.7
2	38.3	5.1	42	71.7	2.7
3	40.1	4.9	43	72.4	2.7
4	41.9	4.5	44	73.1	2.8
5	43.5	4.2	45	73.8	2.8
6	45.0	4.0	46	74.6	2.9
7	46.3	3.8	47	75.3	2.9
8	47.5	3.6	48	76.1	3.0
9	48.7	3.4	49	76.9	3.0
10	49.7	3.3	50	77.8	3.1
11	50.7	3.2	51	78.7	3.2
12	51.6	3.1	52	79.6	3.3
13	52.5	3.0	53	80.6	3.3
14	53.3	2.9	54	81.5	3.3
15	54.1	2.9	55	82.5	3.3
16	54.9	2.8	56	83.5	3.2
17	55.7	2.8	57	84.4	3.1
18	56.4	2.7	58	85.3	2.9
19	57.1	2.7	59	86.0	2.7
20	57.8	2.7	60	86.7	2.4
21	58.5	2.7			
22	59.1	2.6			
23	59.8	2.6			
24	60.4	2.6			
25	61.1	2.6			
26	61.7	2.6			
27	62.3	2.6			
28	62.9	2.6			
29	63.6	2.5			
30	64.2	2.5			
31	64.8	2.5			
32	65.4	2.5			
33	66.0	2.5			
34	66.6	2.5			
35	67.2	2.5			
36	67.9	2.6			
37	68.5	2.6			
38	69.1	2.6			
39	69.8	2.6			

Appendix Table 44: Direct (Raw to Scale) Equipercentile Crosswalk Table - From CES-D to PROMIS Depression (Toolbox Study). Note: Table 43 is recommended.

CES-D Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	34	26	29	0.07
1	34	34	35	0.07
2	34	36	36	0.11
3	39	39	39	0.29
4	43	42	42	0.39
5	44	44	44	0.43
6	45	45	45	0.71
7	47	47	47	0.35
8	48	48	48	0.52
9	49	49	49	0.40
10	50	50	50	0.38
11	51	51	51	0.62
12	52	52	52	0.39
13	53	53	53	0.28
14	53	54	53	0.26
15	55	54	54	0.73
16	56	55	55	0.40
17	56	56	56	0.37
18	57	57	57	0.37
19	58	57	57	0.40
20	58	58	58	0.35
21	59	59	59	0.36
22	59	59	59	0.34
23	59	60	60	0.33
24	60	60	60	0.63
25	61	61	61	0.52
26	62	62	62	0.40
27	62	62	63	0.37
28	63	63	63	0.45
29	64	64	64	0.45
30	64	64	64	0.45
31	65	65	65	0.35
32	65	65	65	0.33
33	65	66	66	0.32
34	66	66	67	0.62
35	66	67	67	1.28
36	67	68	68	1.19
37	68	68	69	0.69
38	69	69	69	1.41

CES-D Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
39	70	70	70	0.69
40	70	70	71	0.65
41	71	71	71	0.66
42	71	71	72	0.59
43	71	72	72	0.56
44	72	72	73	1.41
45	73	73	73	0.67
46	73	73	74	0.56
47	73	74	74	0.44
48	75	75	75	0.71
49	75	76	76	0.75
50	76	77	77	1.62
51	77	77	77	1.58
52	79	78	78	0.82
53	79	79	79	0.82
54	80	80	80	1.27
55	81	81	81	1.17
56	85	83	83	1.00
57	85	85	85	0.61
58	88	86	86	0.61
59	89	88	88	0.61
60	90	90	90	0.61

Appendix Table 45: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From CES-D to PROMIS Depression (Toolbox Study). Note: Table 43 is recommended.

CES-D Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	32	27	10
1	35	34	27
2	37	37	36
3	40	40	40
4	42	42	42
5	44	44	44
6	45	45	46
7	47	46	47
8	48	48	48
9	49	49	49
10	50	50	50
11	51	51	51
12	52	52	52
13	53	53	53
14	53	54	53
15	55	54	54
16	55	55	55
17	56	56	56
18	57	57	56
19	58	57	57
20	58	58	58
21	58	59	58
22	59	59	59
23	60	60	60
24	60	60	60
25	61	61	61
26	62	62	62
27	62	62	62
28	63	63	63
29	64	64	63
30	64	64	64
31	65	65	65
32	65	65	65
33	65	66	66
34	66	66	66
35	67	67	67
36	68	68	68
37	68	68	68
38	69	69	69
39	70	69	69
40	70	70	70

CES-D Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
41	71	70	70
42	71	71	71
43	71	72	72
44	72	72	72
45	73	73	73
46	73	73	74
47	74	74	74
48	75	75	75
49	76	76	76
50	76	76	76
51	78	77	77
52	79	78	78
53	79	79	80
54	80	80	80
55	82	81	81
56	84	82	82
57	85	82	82
58	86	83	83
59	86	84	84
60	86	85	85

Appendix Table 46: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PHQ-9 to PROMIS Depression (Toolbox Study). RECOMMENDED

PHQ-9 Score	PROMIS T-score	SE
0	37.4	6.4
1	42.7	5.3
2	45.9	4.8
3	48.3	4.7
4	50.5	4.3
5	52.5	4.0
6	54.2	3.8
7	55.8	3.7
8	57.2	3.6
9	58.6	3.5
10	59.9	3.4
11	61.1	3.3
12	62.3	3.3
13	63.5	3.2
14	64.7	3.2
15	65.8	3.2
16	66.9	3.2
17	68.0	3.1
18	69.2	3.2
19	70.3	3.2
20	71.5	3.2
21	72.7	3.3
22	74.0	3.4
23	75.3	3.5
24	76.7	3.6
25	78.3	3.7
26	80.0	3.8
27	82.3	3.8

Appendix Table 47: Direct (Raw to Scale) Equipercentile Crosswalk Table - From PHQ-9 to PROMIS Depression (Toolbox Study). Note: Table 46 is recommended.

PHQ-9 Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	34	35	36	0.46
1	44	44	43	0.46
2	47	47	47	0.37
3	49	49	49	0.42
4	51	51	51	0.73
5	53	53	53	0.31
6	55	55	55	0.85
7	57	57	57	0.44
8	58	58	58	0.40
9	59	59	59	0.36
10	60	60	60	0.72
11	62	62	62	0.55
12	63	63	63	0.62
13	64	64	64	0.58
14	65	65	65	0.40
15	65	66	66	0.34
16	66	66	67	0.67
17	67	68	68	1.33
18	69	69	69	1.62
19	70	70	70	0.63
20	71	71	71	0.56
21	72	72	72	1.25
22	73	73	73	0.62
23	73	74	74	0.58
24	75	75	76	0.80
25	77	77	77	1.62
26	79	79	79	0.67
27	84	87	87	1.41

Appendix Table 48: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From PHQ-9 to PROMIS Depression (Toolbox Study). Note: Table 46 is recommended.

PHQ-9 Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	36	37	37
1	44	44	44
2	47	47	47
3	49	49	50
4	51	51	52
5	53	53	53
6	55	55	55
7	57	56	56
8	58	58	57
9	59	59	59
10	60	60	60
11	62	62	61
12	63	63	62
13	64	64	63
14	65	65	65
15	65	66	66
16	66	67	67
17	68	68	68
18	69	69	69
19	70	70	70
20	71	71	71
21	72	72	72
22	73	73	73
23	73	74	74
24	75	75	76
25	77	77	77
26	79	79	79
27	84	83	83

Appendix Table 49: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QOL Anxiety. RECOMMENDED

Neuro-QOL Raw Score	PROMIS Tscore	SE	Neuro-QOL Raw Score	PROMIS Tscore	SE	Neuro-QOL Raw Score	PROMIS Tscore	SE
21	31.5	5.3	61	61.8	1.8	101	83.0	2.6
22	35.3	4.3	62	62.2	1.8	102	84.0	2.6
23	37.4	3.9	63	62.7	1.8	103	84.9	2.5
24	39.2	3.6	64	63.1	1.8	104	85.8	2.4
25	40.7	3.3	65	63.6	1.8	105	86.8	2.2
26	42.0	2.9	66	64.0	1.8			
27	43.1	2.7	67	64.5	1.8			
28	44.1	2.5	68	64.9	1.8			
29	45.0	2.4	69	65.4	1.8			
30	45.9	2.3	70	65.8	1.8			
31	46.6	2.2	71	66.3	1.8			
32	47.3	2.1	72	66.7	1.8			
33	48.0	2.1	73	67.2	1.8			
34	48.7	2.0	74	67.6	1.8			
35	49.3	2.0	75	68.1	1.8			
36	49.9	1.9	76	68.5	1.8			
37	50.4	1.9	77	69.0	1.8			
38	51.0	1.9	78	69.5	1.8			
39	51.5	1.9	79	69.9	1.8			
40	52.0	1.8	80	70.4	1.8			
41	52.6	1.8	81	70.9	1.8			
42	53.1	1.8	82	71.3	1.8			
43	53.6	1.8	83	71.8	1.8			
44	54.1	1.8	84	72.3	1.8			
45	54.5	1.8	85	72.8	1.8			
46	55.0	1.8	86	73.3	1.8			
47	55.5	1.8	87	73.8	1.8			
48	55.9	1.7	88	74.3	1.8			
49	56.4	1.7	89	74.8	1.8			
50	56.9	1.7	90	75.3	1.9			
51	57.3	1.7	91	75.9	1.9			
52	57.8	1.7	92	76.5	1.9			
53	58.2	1.7	93	77.0	2.0			
54	58.7	1.7	94	77.7	2.0			
55	59.1	1.7	95	78.3	2.1			
56	59.6	1.7	96	79.0	2.2			
57	60.0	1.8	97	79.7	2.2			
58	60.5	1.8	98	80.4	2.3			
59	60.9	1.8	99	81.3	2.4			
60	61.4	1.8	100	82.1	2.5			

Appendix Table 50: Direct (Raw to Scale) Equipercentile Crosswalk Table - From Neuro-QOL Anxiety to PROMIS Anxiety. Note: Table 49 is recommended.

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
21	34	29	27	0.07
22	34	34	35	0.07
23	38	37	37	0.20
24	40	39	39	0.15
25	40	40	40	0.13
26	42	41	42	0.26
27	42	43	43	0.21
28	44	44	44	0.27
29	45	45	45	0.36
30	46	46	46	0.19
31	46	47	47	0.20
32	47	48	48	0.34
33	49	48	48	0.12
34	49	49	49	0.12
35	49	49	49	0.12
36	50	50	50	0.34
37	50	50	50	0.32
38	51	51	51	0.42
39	52	52	52	0.16
40	52	52	52	0.15
41	52	52	52	0.14
42	53	53	53	0.25
43	54	54	54	0.26
44	54	54	54	0.24
45	54	54	54	0.21
46	55	55	55	0.23
47	55	55	55	0.22
48	56	56	56	0.25
49	56	56	56	0.26
50	57	57	57	0.21
51	57	57	57	0.18
52	58	58	58	0.33
53	58	58	58	0.26
54	58	59	59	0.27
55	59	59	59	0.67
56	60	60	60	0.27
57	60	60	60	0.24
58	61	61	61	0.18
59	61	61	61	0.15
60	61	61	61	0.16
61	62	62	62	0.14
62	62	62	62	0.14
63	62	62	62	0.13
64	62	63	63	0.11
65	64	63	63	0.29

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
66	64	64	64	0.26
67	65	65	65	0.49
68	66	65	65	0.46
69	66	66	65	0.30
70	66	66	66	0.30
71	66	66	66	0.24
72	66	67	67	0.24
73	67	67	67	0.36
74	68	68	68	1.15
75	69	69	68	0.89
76	70	69	69	0.35
77	70	70	69	0.19
78	70	70	70	0.19
79	70	70	70	0.19
80	70	70	70	0.20
81	70	71	71	0.14
82	70	71	71	0.14
83	72	71	71	0.47
84	72	72	72	0.27
85	72	72	72	0.27
86	72	72	73	1.41
87	72	73	73	1.41
88	73	73	74	1.41
89	74	74	74	1.41
90	75	75	75	1.41
91	76	75	76	1.41
92	76	76	76	1.41
93	76	77	77	1.41
94	78	78	78	0.35
95	78	79	78	0.35
96	79	79	79	0.35
97	85	80	80	0.35
98	85	81	81	0.35
99	85	83	83	0.35
100	85	84	84	0.35
101	85	85	85	0.35
102	85	86	86	0.35
103	85	87	87	0.35
104	85	89	89	0.35
105	85	90	90	0.35

Appendix Table 51: Indirect (Raw to Raw Scale) Equipercentile Crosswalk Table - From Neuro-QOL Anxiety to PROMIS Anxiety. Note: Table 49 is recommended.

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
21	34	34	34
22	35	36	36
23	37	38	38
24	40	39	39
25	40	40	41
26	41	42	42
27	42	43	43
28	44	44	44
29	45	44	45
30	46	46	46
31	46	47	46
32	47	47	47
33	48	48	48
34	49	49	49
35	49	49	49
36	50	50	50
37	50	50	50
38	51	51	51
39	52	52	52
40	52	52	52
41	52	52	52
42	53	53	53
43	54	54	54
44	54	54	54
45	54	54	54
46	55	55	55
47	55	55	55
48	56	56	56
49	56	56	56
50	57	57	57
51	57	57	57
52	58	58	58
53	58	58	58
54	58	59	59
55	59	59	59
56	60	60	60
57	60	60	60
58	61	61	60
59	61	61	61
60	61	61	61
61	62	62	62
62	62	62	62
63	62	62	63

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
64	63	63	63
65	64	64	64
66	64	64	64
67	65	65	64
68	65	65	65
69	66	65	65
70	66	66	66
71	66	66	66
72	66	66	67
73	67	67	67
74	68	68	68
75	68	69	68
76	70	69	69
77	70	70	69
78	70	70	70
79	70	70	70
80	70	71	71
81	71	71	71
82	71	71	71
83	72	72	72
84	72	72	72
85	73	72	73
86	73	73	73
87	73	73	74
88	73	74	74
89	74	74	74
90	75	74	75
91	75	75	75
92	75	76	76
93	76	76	76
94	77	77	77
95	77	78	78
96	79	78	78
97	85	79	79
98	85	80	80
99	85	80	80
100	85	81	81
101	85	82	82
102	85	83	83
103	85	84	84
104	85	85	85
105	85	85	85

Appendix Table 52: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QOL Depression. RECOMMENDED

Neuro-QOL Raw	PROMIS Tscore	SE	Neuro-QOL Raw	PROMIS Tscore	SE	Neuro-QOL Raw	PROMIS Tscore	SE
24	33.4	5.1	64	59.5	1.3	104	73.6	1.3
25	38.0	3.6	65	59.8	1.3	105	74.0	1.4
26	40.1	3.2	66	60.2	1.3	106	74.4	1.4
27	41.7	2.7	67	60.5	1.3	107	74.8	1.4
28	43.1	2.5	68	60.8	1.3	108	75.3	1.4
29	44.1	2.2	69	61.2	1.3	109	75.8	1.4
30	45.1	2.1	70	61.5	1.3	110	76.2	1.5
31	45.9	2.0	71	61.8	1.3	111	76.8	1.5
32	46.6	1.9	72	62.2	1.3	112	77.3	1.6
33	47.3	1.8	73	62.5	1.3	113	77.9	1.6
34	47.9	1.7	74	62.9	1.3	114	78.6	1.7
35	48.5	1.7	75	63.2	1.3	115	79.3	1.8
36	49.0	1.6	76	63.5	1.3	116	80.1	2.0
37	49.5	1.6	77	63.9	1.3	117	81.1	2.1
38	50.0	1.5	78	64.2	1.3	118	82.2	2.3
39	50.5	1.5	79	64.6	1.3	119	83.6	2.5
40	50.9	1.5	80	64.9	1.3	120	85.3	2.5
41	51.4	1.4	81	65.2	1.3			
42	51.8	1.4	82	65.6	1.3			
43	52.2	1.4	83	65.9	1.3			
44	52.6	1.4	84	66.3	1.3			
45	53.0	1.4	85	66.6	1.3			
46	53.4	1.3	86	67.0	1.3			
47	53.7	1.3	87	67.3	1.3			
48	54.1	1.3	88	67.7	1.3			
49	54.4	1.3	89	68.0	1.3			
50	54.8	1.3	90	68.4	1.3			
51	55.1	1.3	91	68.7	1.3			
52	55.5	1.3	92	69.1	1.3			
53	55.8	1.3	93	69.4	1.3			
54	56.2	1.3	94	69.8	1.3			
55	56.5	1.3	95	70.2	1.3			
56	56.8	1.3	96	70.5	1.3			
57	57.2	1.3	97	70.9	1.3			
58	57.5	1.3	98	71.3	1.3			
59	57.8	1.3	99	71.6	1.3			
60	58.2	1.3	100	72.0	1.3			
61	58.5	1.3	101	72.4	1.3			
62	58.8	1.3	102	72.8	1.3			
63	59.2	1.3	103	73.2	1.3			

Appendix Table 53: Direct (Raw to Scale) Equipercentile Crosswalk Table - From Neuro-QOL Depression to PROMIS Depression. Note: Table 52 is recommended.

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
24	35	31	28	0.05
25	35	36	36	0.05
26	40	40	39	0.14
27	42	42	42	0.34
28	44	43	43	0.22
29	44	44	44	0.15
30	45	45	45	0.22
31	46	46	46	0.37
32	47	47	46	0.25
33	47	47	47	0.18
34	48	48	48	0.29
35	48	48	48	0.30
36	49	49	49	0.19
37	49	49	49	0.15
38	50	50	50	0.59
39	51	50	50	0.16
40	51	51	51	0.16
41	51	51	51	0.17
42	52	52	52	0.18
43	52	52	52	0.16
44	52	52	52	0.12
45	53	53	53	0.27
46	54	53	53	0.31
47	54	54	54	0.23
48	54	54	54	0.22
49	55	55	55	0.24
50	55	55	55	0.21
51	55	55	55	0.23
52	56	56	56	0.18
53	56	56	56	0.16
54	56	56	56	0.17
55	56	57	56	0.14
56	57	57	57	0.31
57	57	57	57	0.31
58	58	58	58	0.21
59	58	58	58	0.21
60	58	58	58	0.19
61	59	58	59	0.16
62	59	59	59	0.15
63	59	59	59	0.15
64	59	59	59	0.14
65	59	59	60	0.13

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
66	60	60	60	0.21
67	60	60	60	0.25
68	61	61	61	0.31
69	61	61	61	0.25
70	62	61	61	0.15
71	62	62	62	0.15
72	62	62	62	0.17
73	62	62	62	0.17
74	62	63	63	0.16
75	63	63	63	0.31
76	64	64	64	0.49
77	64	64	64	0.42
78	65	64	64	0.15
79	65	65	65	0.13
80	65	65	65	0.13
81	65	65	65	0.15
82	65	65	65	0.16
83	66	66	66	0.30
84	66	66	66	0.22
85	66	67	67	0.18
86	67	67	67	0.51
87	68	68	67	0.28
88	68	68	68	0.25
89	68	68	68	0.16
90	68	69	69	0.19
91	70	69	69	0.47
92	70	70	70	0.42
93	70	70	70	0.42
94	70	70	70	0.71
95	71	71	71	0.35
96	71	71	71	0.38
97	72	71	71	0.47
98	72	72	72	0.47
99	72	72	72	0.38
100	72	72	72	0.38
101	72	72	72	0.38
102	72	72	73	0.47
103	72	73	73	2.45
104	74	73	73	0.71
105	74	74	74	0.35
106	74	74	74	0.35
107	74	74	75	0.35
108	74	75	75	0.35
109	76	76	76	0.71
110	77	77	77	0.02
111	78	78	78	0.02

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
112	78	79	79	1.41
113	78	79	80	1.41
114	78	80	80	1.41
115	79	81	81	1.41
116	80	82	83	0.01
117	81	84	85	0.01
118	88	86	86	0.02
119	89	88	88	0.02
120	90	90	90	0.02

Appendix Table 54: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From Neuro-QOL Depression to PROMIS Depression. Note: Table 52 is recommended.

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
24	35	35	35
25	37	37	37
26	40	40	40
27	42	41	41
28	43	43	43
29	44	44	44
30	45	45	45
31	46	46	46
32	47	47	46
33	47	47	47
34	48	48	48
35	48	48	48
36	49	49	49
37	49	49	49
38	50	50	50
39	50	50	50
40	51	51	51
41	51	51	51
42	52	52	52
43	52	52	52
44	52	52	53
45	53	53	53
46	54	54	53
47	54	54	54
48	54	54	54
49	54	55	54
50	55	55	55
51	55	55	55
52	56	56	56
53	56	56	56
54	56	56	56
55	56	57	57
56	57	57	57
57	58	57	57
58	58	58	58
59	58	58	58
60	58	58	58
61	58	58	58
62	59	59	59
63	59	59	59
64	59	59	59
65	59	59	60

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
66	60	60	60
67	60	60	60
68	61	61	61
69	61	61	61
70	62	62	61
71	62	62	62
72	62	62	62
73	62	62	62
74	63	63	63
75	63	63	63
76	64	64	64
77	64	64	64
78	65	64	64
79	65	65	64
80	65	65	65
81	65	65	65
82	65	65	65
83	65	66	66
84	66	66	66
85	66	66	67
86	67	67	67
87	68	68	67
88	68	68	68
89	68	68	68
90	68	69	69
91	70	69	69
92	70	70	70
93	70	70	70
94	70	70	70
95	71	71	71
96	71	71	71
97	71	71	71
98	72	72	72
99	72	72	72
100	72	72	72
101	72	72	72
102	72	73	73
103	72	73	73
104	74	74	73
105	74	74	74
106	74	74	74
107	74	74	74
108	74	74	74
109	76	75	75
110	77	77	77

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
111	78	78	78
112	79	80	80
113	79	80	81
114	79	80	82
115	79	81	84
116	80	82	84
117	81	83	84
118	81	83	84
119	82	84	84
120	84	84	84

Appendix Table 55: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QOL Mobility. RECOMMENDED.

Neuro-QOL Raw	PROMIS Tscore	SE	Neuro-QOL Raw	PROMIS Tscore	SE
19	12.4	1.7	59	33.6	1.3
20	13.4	1.9	60	33.9	1.3
21	14.5	2.1	61	34.3	1.3
22	15.5	2.1	62	34.6	1.3
23	16.5	2.0	63	35.0	1.4
24	17.4	1.9	64	35.4	1.4
25	18.2	1.9	65	35.7	1.4
26	18.9	1.8	66	36.1	1.4
27	19.6	1.7	67	36.4	1.4
28	20.3	1.7	68	36.8	1.4
29	20.9	1.6	69	37.2	1.4
30	21.4	1.6	70	37.6	1.4
31	22.0	1.6	71	38.0	1.4
32	22.5	1.6	72	38.3	1.4
33	23.0	1.5	73	38.7	1.4
34	23.5	1.5	74	39.1	1.4
35	24.0	1.5	75	39.6	1.4
36	24.5	1.5	76	40.0	1.5
37	24.9	1.5	77	40.4	1.5
38	25.4	1.5	78	40.8	1.5
39	25.8	1.5	79	41.3	1.5
40	26.3	1.4	80	41.8	1.5
41	26.7	1.4	81	42.3	1.6
42	27.1	1.4	82	42.8	1.6
43	27.5	1.4	83	43.3	1.6
44	27.9	1.4	84	43.9	1.7
45	28.3	1.4	85	44.5	1.7
46	28.7	1.4	86	45.1	1.8
47	29.1	1.4	87	45.8	1.9
48	29.5	1.4	88	46.5	2.0
49	29.9	1.4	89	47.4	2.1
50	30.3	1.4	90	48.4	2.3
51	30.7	1.4	91	49.5	2.6
52	31.0	1.4	92	50.9	2.8
53	31.4	1.4	93	52.6	3.1
54	31.8	1.3	94	55.2	3.7
55	32.1	1.3	95	61.3	5.9
56	32.5	1.3			
57	32.8	1.3			
58	33.2	1.3			

Appendix Table 56: Direct (Raw to Scale) Equipercentile Crosswalk Table - From Neuro-QOL Mobility to PROMIS Physical Function. Note: Table 55 is recommended.

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
19	14	10	10	0.35
20	14	11	11	0.35
21	14	12	12	0.35
22	14	13	13	0.35
23	14	14	14	0.35
24	15	15	15	1.41
25	18	16	16	1.41
26	18	17	17	0.71
27	18	18	18	0.71
28	20	19	19	0.71
29	20	20	20	0.61
30	20	21	21	0.61
31	22	21	21	0.35
32	22	22	22	0.41
33	22	22	22	0.72
34	23	23	23	0.62
35	23	23	23	0.66
36	24	24	24	0.45
37	24	24	24	0.40
38	24	25	25	0.78
39	25	25	25	0.61
40	26	26	26	0.33
41	27	27	27	0.65
42	27	27	27	0.51
43	28	28	28	0.51
44	28	28	28	0.44
45	29	29	29	0.40
46	29	29	29	0.45
47	30	29	29	0.15
48	30	30	30	0.15
49	30	30	30	0.14
50	30	30	30	0.14
51	30	30	30	0.15
52	30	31	31	0.16
53	31	31	31	0.35
54	31	31	31	0.35
55	32	32	32	0.35
56	32	32	32	0.33
57	33	33	33	0.51
58	33	33	33	0.47
59	34	34	34	0.14
60	34	34	34	0.12

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
61	34	34	34	0.14
62	34	35	35	0.15
63	35	35	35	0.51
64	36	36	36	0.28
65	36	36	36	0.28
66	37	37	36	0.22
67	37	37	37	0.20
68	37	37	37	0.20
69	38	38	38	0.11
70	38	38	38	0.10
71	38	38	38	0.11
72	38	38	38	0.12
73	39	39	39	0.25
74	39	39	39	0.24
75	40	40	40	0.23
76	40	40	40	0.21
77	40	40	40	0.19
78	41	41	41	0.18
79	41	41	41	0.16
80	41	42	42	0.16
81	43	43	43	0.12
82	43	43	43	0.12
83	43	43	43	0.14
84	44	44	44	0.13
85	44	44	44	0.14
86	46	45	45	0.12
87	46	46	46	0.11
88	46	46	47	0.10
89	49	48	48	0.07
90	49	49	49	0.07
91	49	49	49	0.06
92	49	50	50	0.07
93	57	57	56	0.03
94	57	57	57	0.03
95	57	66	67	0.03

Appendix Table 57: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From Neuro-QOL Mobility to PROMIS Physical Function. Note: Table 55 is recommended.

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
19	14	14	14
20	14	14	15
21	14	15	15
22	14	16	16
23	14	17	17
24	16	18	18
25	18	18	18
26	19	19	19
27	19	19	20
28	20	20	20
29	20	20	20
30	20	21	21
31	22	22	22
32	22	22	22
33	23	23	23
34	23	23	23
35	24	24	24
36	24	24	24
37	24	25	25
38	25	25	25
39	26	26	26
40	26	26	26
41	27	27	27
42	27	27	27
43	28	28	28
44	28	28	28
45	29	28	28
46	29	29	29
47	29	29	29
48	30	30	30
49	30	30	30
50	30	30	30
51	30	31	31
52	31	31	31
53	31	31	31
54	31	32	32
55	32	32	32
56	32	32	32
57	32	33	33
58	33	33	33
59	33	33	34
60	34	34	34
61	34	34	34
62	34	34	35
63	35	35	35

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
64	35	35	35
65	36	36	36
66	36	36	36
67	37	37	37
68	37	37	37
69	37	37	37
70	38	38	38
71	38	38	38
72	39	39	38
73	39	39	39
74	40	39	39
75	40	40	40
76	40	40	40
77	40	41	40
78	41	41	41
79	42	42	41
80	42	42	42
81	42	42	42
82	43	43	43
83	43	43	43
84	44	44	44
85	45	44	44
86	45	45	45
87	46	46	46
88	47	46	46
89	47	47	47
90	48	48	48
91	50	49	49
92	52	51	51
93	53	53	54
94	56	56	57
95	57	57	57

Appendix Table 58: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QOL Upper Extremity. RECOMMENDED

Neuro-QOL Raw	PROMIS Tscore	SE	Neuro-QOL Raw	PROMIS Tscore	SE
20	10.7	0.7	60	24.6	1.6
21	10.8	0.8	61	24.9	1.6
22	10.9	0.8	62	25.3	1.6
23	11.1	0.9	63	25.7	1.6
24	11.2	1.0	64	26.0	1.6
25	11.4	1.1	65	26.4	1.6
26	11.6	1.2	66	26.7	1.6
27	11.8	1.3	67	27.1	1.7
28	12.1	1.4	68	27.5	1.7
29	12.5	1.5	69	27.9	1.7
30	12.8	1.6	70	28.2	1.7
31	13.2	1.7	71	28.6	1.7
32	13.6	1.7	72	29.0	1.7
33	14.1	1.8	73	29.4	1.7
34	14.5	1.8	74	29.8	1.7
35	15.0	1.8	75	30.2	1.7
36	15.4	1.8	76	30.6	1.7
37	15.9	1.8	77	31.0	1.7
38	16.3	1.8	78	31.4	1.7
39	16.7	1.8	79	31.8	1.8
40	17.1	1.7	80	32.3	1.8
41	17.5	1.7	81	32.7	1.8
42	17.9	1.7	82	33.2	1.8
43	18.3	1.7	83	33.7	1.8
44	18.7	1.7	84	34.2	1.9
45	19.1	1.7	85	34.7	1.9
46	19.5	1.7	86	35.2	1.9
47	19.9	1.7	87	35.8	2.0
48	20.3	1.7	88	36.4	2.0
49	20.6	1.7	89	37.0	2.1
50	21.0	1.7	90	37.7	2.2
51	21.4	1.7	91	38.4	2.3
52	21.7	1.6	92	39.2	2.4
53	22.1	1.6	93	40.1	2.6
54	22.4	1.6	94	41.1	2.8
55	22.8	1.6	95	42.3	3.1
56	23.2	1.6	96	43.9	3.7
57	23.5	1.6	97	45.4	3.9
58	23.9	1.6	98	47.4	4.3
59	24.2	1.6	99	50.0	4.7
60	24.6	1.6	100	57.4	6.9

Appendix Table 59: Direct (Raw to Scale) Equipercentile Crosswalk Table - From Neuro-QOL Upper Extremity to PROMIS Physical Function. Note: Table 58 is recommended.

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
20	12	10	10	0.13
21	12	11	11	0.13
22	12	11	11	0.13
23	12	12	12	0.14
24	12	12	12	0.14
25	12	12	12	0.13
26	12	12	12	0.13
27	12	12	13	0.12
28	12	13	13	0.14
29	12	13	13	0.47
30	13	13	13	0.47
31	13	13	13	0.54
32	14	14	14	1.27
33	15	15	15	0.47
34	15	15	15	0.26
35	15	15	15	0.28
36	16	16	16	1.41
37	17	16	16	0.67
38	17	16	16	0.47
39	17	17	16	0.47
40	17	17	17	0.47
41	17	17	17	0.54
42	17	17	17	0.54
43	18	17	17	2.00
44	18	17	18	2.00
45	18	18	18	2.00
46	18	18	18	2.00
47	19	19	19	0.86
48	20	20	19	0.26
49	20	20	20	0.28
50	20	20	20	0.30
51	20	20	20	0.30
52	20	21	21	0.60
53	21	21	21	0.52
54	21	21	21	0.52
55	22	22	22	0.60
56	23	23	23	0.71
57	23	23	23	0.72
58	24	24	24	0.47
59	24	24	24	0.49
60	25	25	24	0.23
61	25	25	25	0.23
62	25	25	25	0.23
63	25	25	25	0.20
64	26	26	26	0.15
65	26	26	26	0.15

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing) .0.3	Equipercentile Equivalents with Postsmoothing (More Smoothing) 1.0	Standard Error of Equating (SEE)
66	26	26	26	0.16
67	26	26	26	0.16
68	26	27	27	0.17
69	27	27	27	0.50
70	28	28	28	0.19
71	28	28	28	0.20
72	29	29	29	0.49
73	29	29	29	0.46
74	30	30	30	0.22
75	30	30	30	0.21
76	30	31	30	0.18
77	31	31	31	0.39
78	32	31	31	0.14
79	32	32	32	0.14
80	32	32	32	0.15
81	32	32	32	0.14
82	33	33	33	0.36
83	33	33	33	0.33
84	34	34	34	0.38
85	35	35	35	0.20
86	35	35	35	0.17
87	36	36	36	0.19
88	36	36	36	0.24
89	37	37	37	0.24
90	37	38	38	0.21
91	39	39	39	0.14
92	39	39	39	0.15
93	40	40	40	0.16
94	43	42	42	0.13
95	43	43	43	0.12
96	45	44	44	0.09
97	45	45	45	0.09
98	55	54	54	0.02
99	55	55	55	0.02
100	55	60	60	0.02

Appendix Table 60: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From Neuro-QOL Upper Extremity to PROMIS Physical Function. Note: Table 58 is recommended.

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
20	12	12	12
21	12	12	12
22	12	12	12
23	12	12	12
24	12	12	12
25	12	12	12
26	12	12	12
27	12	12	12
28	12	12	13
29	13	13	13
30	13	13	13
31	13	14	14
32	14	14	14
33	15	15	14
34	15	15	15
35	16	16	15
36	16	16	16
37	17	16	16
38	17	17	16
39	17	17	17
40	17	17	17
41	17	17	17
42	18	18	17
43	18	18	18
44	18	18	18
45	18	18	18
46	18	19	19
47	19	19	19
48	20	20	20
49	20	20	20
50	20	20	20
51	20	21	21
52	21	21	21
53	21	21	21
54	21	22	22
55	22	22	22
56	23	22	22
57	23	23	23
58	23	24	23
59	24	24	24
60	25	24	24
61	25	25	25
62	25	25	25
63	25	25	25
64	26	26	26
65	26	26	26

Neuro-QOL Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
66	26	26	26
67	26	26	26
68	27	27	27
69	27	27	27
70	27	27	28
71	28	28	28
72	29	28	28
73	29	29	29
74	29	29	29
75	30	30	30
76	30	30	30
77	31	31	31
78	31	31	31
79	32	32	32
80	32	32	32
81	32	33	33
82	33	33	33
83	33	34	34
84	34	34	34
85	35	34	34
86	35	35	35
87	36	36	36
88	36	36	36
89	37	37	37
90	38	38	38
91	39	39	39
92	40	40	40
93	41	41	41
94	42	42	42
95	43	43	43
96	44	44	44
97	46	46	46
98	49	48	48
99	51	51	51
100	55	55	55