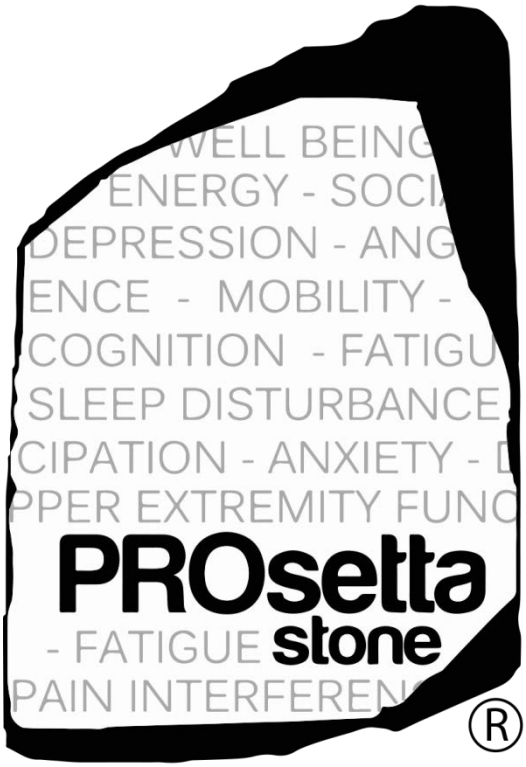


PROSETTA STONE® ANALYSIS REPORT

A ROSETTA STONE FOR PATIENT REPORTED OUTCOMES

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

DEPARTMENT OF MEDICAL SOCIAL SCIENCES
FEINBERG SCHOOL OF MEDICINE
NORTHWESTERN UNIVERSITY

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

Table of Contents

PRO Rosetta Stone (*PROsetta Stone®*) Analysis

1. Introduction

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

2. The PRO Rosetta Stone Project

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

2.1. Patient-Reported Outcomes Measurement Information System (PROMIS)

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

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

The PROMIS Wave I data collection design included a number of widely accepted “legacy” measures. The legacy measures used for validation evidence included Buss-Perry Aggression Questionnaire (BPAQ), Center for Epidemiological Studies Depression Scale (CES-D), Mood and Anxiety Symptom Questionnaire (MASQ), Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F), Brief Pain Inventory (BPI), and SF-36. Furthermore, included within each of the PROMIS banks were items from several other existing measures. Depending on the nature and strength of relationship between the measures, various linking procedures can be used to allow for cross-walking of scores. (Most of the linking reports based on the PROMIS Wave 1 dataset are included in Volume 1)(Choi et al., 2012).

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

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

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

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

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

2.4. PROsetta Stone Data Collection

The National Institutes of Health/National Cancer Institute supported three waves of data collection as part of the PROseta Stone project. The specific aim of each data collection was to administer a range of PROMIS instruments along with legacy measures, following a single sample design (Kolen & Brennan, 2004). For adults (Waves 1 and 2), the assessed (sub)domains comprised negative affect (anger, anxiety, and depression), fatigue, cognitive

function, global health, pain interference, physical function, satisfaction with social relationships and activities, sleep disturbance, sleep-related impairment, positive affect and well-being. For children (Wave 3), the following (sub)domains were assessed: anxiety, depression, fatigue, cognitive function, peer relationships, and physical function. The PROsetta Stone data collection allowed investigators to make links to commonly used instruments not administered in PROMIS, Neuro-QoL, and NIH Toolbox studies.

3. Legacy Instruments

The following instruments are widely accepted "legacy" measures that were linked to PROMIS instruments. Some of these legacy measures were used as part of the initial validation work for PROMIS and NIH Toolbox, or administered as part of this PROsetta Stone project. Data were collected on a minimum of 400 respondents (for stable item parameter estimation) along with at least one other conceptually similar scale or bank. (See Table 5.1).

3.1. Functional Assessment of Cancer Therapy-Cognitive Function (FACT-Cog, Version 3)

The Functional Assessment of Cancer Therapy-Cognitive Function (FACT-Cog, Version 3) is a 37-item self-report questionnaire to assess cognitive function in cancer patients before, during, and after chemotherapy, specifically their memory, attention, concentration, language and thinking abilities. The FACT-Cog consists of four subscales. In the Perceived Cognitive Impairments and the Comments from Others subscales the patient is asked to indicate how often each situation occurred during the last seven days, using a 5-point Likert-type scale ("from 0 "Never" to 4 "Several times a day"). An intensity 5-point Likert-type scale (from 0 "Not at all" to 4 "Very much") is used to rate Perceived Cognitive Abilities and the Impact on Quality of Life. For all subscales, a higher score represents better cognitive functioning or quality of life. Scoring includes calculation of the four subscales: Perceived Cognitive Impairments (20 items; score range: 0-72), Impact on QOL (4 items; score range: 0-16), Comments from Others (4 items; score range: 0-16) and Perceived Cognitive Abilities (9 items; score range: 0-28).

3.2. Pediatric Perceived Cognitive Function Item Bank (Peds PCF)

The Pediatric Perceived Cognitive Function Item Bank (Ped PCF) consists of 43 items measuring children's cognitive behaviors. Both parent-reported and child-reported versions are available. The Ped PCF was initially designed for children with cancer who receive neurotoxicity treatments and for other populations of children and adolescents at risk for neurocognitive impairment. The Ped PCF has satisfactory psychometric properties, as evaluated using both classical test theory and IRT approaches, in use with the US general population (Lai et al, 2011)and with children with cancer. (Lai et al., In Press) It produces reliable scores that can discriminate between children with (versus without) significant symptoms of attention, social,

and thought problems as well as between children with brain tumors versus those with other types of cancer. US general population-based norms are available to serve as a reference. This measure uses two 5-point rating scales: One is frequency related: (“none of the time” to “all of the time”) and one is intensity related (“not at all” to “very much”). For context, a 4-week timeframe is used. A 7-item short form and a computer adaptive test (CAT) version of the item bank are available.

3.3. Hospital Anxiety and Depression Scale (HADS)

The Hospital Anxiety and Depression Scale (HADS) is a 14 item instrument developed by Zigmond and Snaith (Zigmond & Snaith, 1983) to determine levels of anxiety and depression in patients in hospital outpatient clinics. There are seven items each for anxiety and depression each scored from 0 to 3 for a possible total of 0 to 21 for either anxiety or depression. A score of 0 to 7 is considered a non-case, 8 to 10 is considered a borderline case, and 11 or greater is considered a case.

3.4. Positive Affect Negative Affect Schedule (PANAS)

The Positive and Negative Affect Schedule (PANAS) is a 20-item instrument that comprises two scales measuring positive and negative affect, which are described as important dimensions of mood .(Watson, Clark, & Tellegen, 1988) The instrument consists of a number of words that describe different feelings and emotions. Any of a number of time instructions can be given at the researcher’s discretion. The respondent is asked to read each word (item) and then mark the appropriate answer in the space next to that word. Each item is rated on a five-point scale with 1 for very slightly or not at all, 2 for a little, 3 for moderately, 4 for quite a bit, and 5 for extremely. Positive affect is represented by the words enthusiastic, interested, determined, excited, inspired, alert, active, strong, proud, and attentive while negative affect is represented by scared, afraid, upset, distressed, jittery, nervous, ashamed, guilty, irritable, and hostile. For positive affect (PA), a higher score indicates more positive affect, or the extent to which the individual feels enthusiastic, active, and alert. High PA is a state of high energy, full concentration, and pleasurable engagement, whereas low PA is characterized by sadness and lethargy. For negative affect (NA), a higher score indicates more negative affect, or the extent to which the individual feels general subjective distress and ’unpleasureable’ engagement that subsumes a variety of aversive mood states, including anger, contempt, disgust, guilt, fear, and nervousness. Low NA is a state of calmness and serenity.

3.5. Beck Depression Inventory, second edition (BDI-II)

The Beck Depressive Inventory (BDI) is a 21 item instrument for measuring the severity of depression with each answer being scored on a scale value of 0 to 3. The cutoffs used are 0 to 13 for minimal depression, 14 to 19 for mild depression, 20 to 28 for moderate depression, and 29 to 63 for severe depression. Higher total scores indicate more severe depressive symptoms.

Three versions have been developed. The original BDI (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) was revised beginning in 1971 (BDI-1A) (Beck & Steer, 1993), which eliminated the alternative wordings for the same symptoms and the double negatives in the original version. The number of alternatives per item was reduced to three, and the wording was changed for 15 items. Several pilot versions of the BDI-IA were tested, and Beck copyrighted the final version in 1978. With the release the American Psychiatric Association's (1994) *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.) (DSM-IV), he upgraded the amended version to the Beck Depression Inventory, second edition (BDI-II), (Beck, Steer, & Brown, 1996) and Beck, Steer, Ball, and Ranieri) (Beck, Steer, Ball, & Ranieri, 1996). He added symptoms that addressed DSM-IV criteria for major depression disorders, such as Agitation, Concentration, Difficulty, and Worthlessness. The BDI symptoms of Weight Loss, Body Image Change, and Somatic Preoccupation were dropped from the BDI-II because a series of psychometric analyses demonstrated these symptoms were less indicative of the overall severity of depression in 1996 than these same items had been in 1961. The majority of the retained BDI-II items were rewritten for clarity.

3.6. Kessler Psychological Distress Scale (K6)

The Kessler Psychological Distress Scale (K6) is a simple measure of psychological distress which involves 6 questions about a person's emotional state. 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. Each question is scored from 0 (None of the time) to 4 (All of the time). Scores of the 6 questions are then summed, yielding a minimum score of 0 and a maximum score of 24. Low scores indicate low levels of psychological distress and high scores indicate high levels of psychological distress. The K10 and the K6 scales are administered in Australia using an alternate scoring system based on responses of "1-5" versus the "0-4" system presented here. This alternate system results in a score range of 6-30 for the K6 and 10-50 for the K10. The optimal cut point on the K6 for this system is 6-18 versus 19+. The scoring rules are provided separately for each country to convert K6 scores into predicted probabilities of serious mental illness (Kessler, Green, Adler, & et al., 2010).

3.7. Patient Health Questionnaire-2 (PHQ-2)

The Patient Health Questionnaire-2 (PHQ-2) comprises the first two items of the nine-item PHQ depression module or PHQ-9. The PHQ-2 inquires about the frequency of depressed mood and anhedonia over the past two weeks and is used as a screener rather than to diagnose a depressive disorder or to measure depression severity. Further evaluation with the PHQ-9 is recommended for patients who screen positive in the PHQ-2 assessment. A PHQ-2 score

ranges from 0 to 6, with each item scoring as 0 ("not at all") to 3 ("nearly every day"). A score of 3 is considered the optimal cutoff point for screening purposes (Kroenke, Spitzer, & Williams, 2003).

3.8. Veterans RAND 12 Item Health Survey (VR-12)

The Veterans RAND 12 Item Health Survey (VR-12) is a brief, generic, multi-use, self-administered health survey comprised of 12 items (Kazis et al., 2004; Selim et al., 2009). The instrument is primarily used to measure health related quality of life, to estimate disease burden and to evaluate disease-specific benchmarks with other populations. The 12 items in the questionnaire correspond to eight principal physical and mental health domains including general health perceptions; physical functioning; role limitations due to physical and emotional problems; bodily pain; energy-fatigue, social functioning and mental health. The 12 items are summarized into two scores, a Physical Health Summary Measure (PCS-physical component score) and a Mental Health Summary Measure (MCS-mental component score). These provide an important contrast between physical and psychological health status.

3.9. 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.

3.10. Pittsburgh Sleep Quality Index (PSQI)

The Pittsburgh Sleep Quality Index (PSQI) is a self-rated questionnaire which assesses sleep quality and disturbances over a 1-month time interval. Nineteen individual items generate seven "component" scores: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. The sum of scores for these seven components yields one global score. A global PSQI score greater than 5 yields a diagnostic sensitivity in distinguishing good and poor sleepers. The properties of the PSQI suggest that it is useful both in psychiatric clinical practice and research activities (Buysse, Reynolds, Monk, Berman, & Kupfer, 1989).

4. Linking Methods

PROMIS full-bank administration allows for single group linking. This linking method is used when two or more measures are administered to the same group of people. For example, two PROMIS banks (Anxiety and Depression) and three legacy measures (MASQ, CES-D, and SF-36/MH) were administered to a sample of 925 people. The order of measures was randomized so as to minimize potential order effects. The original purpose of the full-bank administration study was to establish initial validity evidence (e.g., validity coefficients), not to establish linking relationships. Some of the measures revealed severely skewed score distributions in the full-bank administration sample and the sample size was relatively small, which might be limiting factors when it comes to determining the linking method. Another potential issue is related to how the non-PROMIS measures are scored and reported. For example, all SF-36 subscales are scored using a proprietary scoring algorithm and reported as normed scores (0 to 100). Others are scored and reported using simple raw summed scores. All PROMIS measures are scored using the final re-centered item response theory (IRT) item parameters and transformed to the T-score metric (mean=50, SD=10).

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

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

4.1. IRT Linking

One of the objectives of the current linking analysis is to determine whether or not the non-PROMIS measures can be added to their respective PROMIS item bank without significantly altering the underlying trait being measured. The rationale is twofold: (1) the augmented

PROMIS item banks might provide more robust coverage both in terms of content and difficulty; and (2) calibrating the non-PROMIS measures on the corresponding PROMIS item bank scale might facilitate subsequent linking analyses. At least, two IRT linking approaches are applicable under the current study design; (1) linking separate calibrations through the Stocking-Lord method and (2) fixed parameter calibration.

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

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

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

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

4.2. Equipercentile Linking

The IRT Linking procedures described above are permissible only if the traits being measured are not significantly altered by aggregating items from multiple measures. One potential issue might be creating multidimensionality as a result of aggregating items measuring different traits. For two scales that measure distinct but highly related traits, predicting scores on one scale from those of the other has been used frequently. Concordance tables between PROMIS and non-PROMIS measures can be constructed using equipercentile equating (Kolen & Brennan, 2004; Lord, 1982) when there is insufficient empirical evidence that the instruments measure the same construct. An equipercentile method estimates a nonlinear linking relationship using percentile rank distributions of the two linking measures. The equipercentile linking method can be used in conjunction with a presmoothing method such as the loglinear model (Hanson, Zeng, & Colton, 1994). The frequency distributions are first smoothed using the loglinear model and then equipercentile linking is conducted based on the smoothed frequency distributions of the two measures. Smoothing can also be done at the backend on equipercentile equivalents and is called postsmoothing(Brennan, 2004; Kolen & Brennan, 2004)The cubic-spline smoothing algorithm (Reinsch, 1967) is used in the LEGS program(Brennan, 2004). Smoothing is intended to reduce sampling error involved in the linking process. A successful linking procedure will provide a conversion (crosswalk) table, in which, for example, raw summed scores on the PHQ-9 measure are transformed to the T-score equivalents of the PROMIS Depression measure.

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

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

4.3. Assumptions and Planned Linking

In Section 5 of this PROsetta Stone report, we present the results of a large number of linking studies using a combination of newly collected and secondary data sets. In most cases, we have applied all three linking methods described in sections 4.1 and 4.2. Our purpose is to provide the maximum amount of useful information. However, the suitability of these methods depends upon the meeting of various linking assumptions. These assumptions require that the two instruments to be linked measure the same construct, show a high correlation, and are relatively invariant in subpopulation differences(Dorans, 2007). The degree to which these assumptions are met varies across linking studies. Given that different researchers may interpret these requirements differently, we have taken a liberal approach for inclusion of linkages in this book. Nevertheless, we recommend that researchers diagnostically review the classical psychometrics and CFA results in light of these assumptions prior to any application of the cross-walk charts or legacy parameters to their own data.

Having investigated a large number of possible links between PROMIS and legacy measures, we did apply a few minimal exclusion rules before linking. We generally did not proceed with planned linking when the raw score correlation between two instruments was less than .70. Table 4.3.1 shows the pairs of adult instruments we planned to link, but did not complete because this requirements was not met.

Table 4.3.1 Planned Adult Instrument Pairs not Meeting Linking Criteria

Planned Instrument Linking Pair	Reason for not Linking
PROMIS Applied Cognitive Abilities and Neuro-QoL Executive Function	Correlation = .66
PROMIS Satisfaction with Social Relationships and Activities (v2) and Neuro-QoL Social Relationships and Activities	Correlation = .61
PROMIS Satisfaction with Social Relationships and Activities (v2) and FACT Social Well-Being	Correlation = .63
PROMIS Ability to Participate in Social Relationships and Activities (v2) and Neuro-QoL Ability to Participate in Social Relationships and Activities	Correlation = .63
PROMIS Ability to Participate in Social Relationships and Activities (v2) and FACT Social Well-Being	Correlation = .41
PROMIS Sleep Disturbance and Epworth Sleepiness Scale	Correlation = .24
PROMIS Sleep-related Impairment and Epworth Sleepiness Scale	Correlation = .46

In other cases, we identified two measures apparently suitable for linking, but were unable to obtain the sufficient data. That is, we typically sought datasets of sufficient size (N > 400) such that IRT linking was feasible. Other reasons for not linking included having only computer

adaptive test (CAT) administration of PROMIS measures and lacking a single sample in which both instruments were administered. Table 4.3.2 shows instruments pairs we planned to link, but were unable to because the required data was unavailable.

Table 4.3.2. Planned Adult Instrument Pairs not Linked – Data Not Available

Planned Instrument Linking Pair	Reason for Not Linking
PROMIS Depression and Apathy Evaluation Scale	Available data limited to age > 50
PROMIS Physical Function and KCCQ Physical Limitation	Sample size less than < 400
PROMIS Physical Function and PAQ	Sample size less than < 400
PROMIS Sleep Disturbance and MOS Sleep Questionnaire	CAT administration of PROMIS
PROMIS Sleep Impairment and MOS Sleep Questionnaire	CAT administration of PROMIS

5. Linking Results

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

Table 5.1. Linking by Reference Instrument

Section	PROMIS Instrument	Instrument to Link	Study
5.1.	PROMIS Cognitive Function-Abilities	FACT-Cog Perceived Cog. Abilities	PROsetta W2
5.2.	PROMIS Cognitive Function v2.0	FACT-Cog Perceived Cog Perceived Cog. Impairment	PROsetta W2
5.3.	PROMIS Cognitive Function v2.0	Neuro-QoL Applied Cog. General Concerns *	PROsetta W2
5.4	PROMIS Cognitive Function v2.0	Peds PCF Short Form	PROsetta W2
5.5	PROMIS Anxiety	HADS	PROsetta W1
5.6	PROMIS Anxiety	PANAS	PROsetta W1
5.7	PROMIS Depression	BDI-II	PROsetta W1
5.8	PROMIS Depression	K6	NIH Toolbox CV
5.9	PROMIS Depression	PANAS	NIH Toolbox CV
5.10	PROMIS Depression	PHQ-2	NIH Toolbox CV
5.11	PROMIS Fatigue	Neuro-QoL Fatigue	PROsetta W1
5.12	PROMIS Global Health - Mental	VR-12 – Mental Component (Sums)	PROsetta W2
5.13	PROMIS Global Health - Mental	VR-12 – Mental Component (Algorithmic)	PROsetta W2
5.14	PROMIS Global Health - Physical	VR-12 – Physical Component (Sums)	PROsetta W2
5.15	PROMIS Global Health - Physical	VR-12 – Physical Component (Algorithmic)	PROsetta W2
5.16	PROMIS Pain Interference	SF-36/BP	PROMIS W1
5.17	PROMIS Sleep Disturbance	Neuro-QoL Sleep Disturbance	PROsetta W2
5.18	PROMIS Sleep Disturbance	PROMIS Sleep-related Impairment	PROsetta W2
5.19	PROMIS Sleep Disturbance	PSQI	PROMIS W1
5.20	PROMIS Sleep-related Impairment	Neuro-QoL Sleep Disturbance	PROsetta W2
5.21	PROMIS Sleep-related Impairment	PSQI	PROMIS W1
5.22	PROMIS Satisfaction w/ Social Roles & Activities v2.0	PROMIS Satisfaction w/ Participation in Discretionary Social Activities v1.0	PROsetta W2
5.23	PROMIS Satisfaction w/ Social Roles & Activities v2.0	PROMIS Satisfaction w/ Participation in Social Roles v1.0	PROsetta W2

Section	Neuro-QoL Instrument	Instrument to Link	Study
5.24	Neuro-QoL Positive Affect & Well-being	NIH Toolbox Life Satisfaction	PROsetta W2
5.25	Neuro-QoL Positive Affect & Well-being	NIH Toolbox Meaning	PROsetta W2
5.26	Neuro-QoL Positive Affect & Well-being	NIH Toolbox Positive Affect	PROsetta W2
Section	**PROMIS Instrument**	**Instrument to Link**	**Study**
5.27	PROMIS Cognitive Function v2.0	Neuro-QoL Cognitive Function v2.0	PROsetta W2

** In 2014, the two Neuro-QoL Applied Cognition banks -- General Concerns and Executive Function – were merged into a single bank called Neuro-QoL Cognitive Function. This new bank was linked via common items to the PROMIS Cognitive Function v2.0 bank, so that the T-scores from either instrument are on the same metric. See Report 5.27 for details on the link with Neuro-QoL Cognitive Function v2.0.*

5.1. PROMIS Cognitive Function-Abilities and FACT-Cog Perceived Cognitive Abilities

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Cognition, namely the PROMIS Cognitive Function - Abilities item bank (31 items) and FACT-Cog Perceived Cognitive Abilities (9 items). PROMIS Cognitive Function - Abilities was scaled such that higher scores represent higher levels of cognition. 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. Our sample consisted of 1,009 participants (N = 1,005 for participants with complete responses).

5.1.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 155 for PROMIS Cog Abilities and 45 for FACT Cog Abilities. Figures 5.1.1 and 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 Cog Abilities and FACT Cog Abilities was 0.87. The disattenuated (corrected for unreliabilities) correlation between PROMIS Cog Abilities and FACT Cog Abilities was 0.9. The correlations between the combined score and the measures were 1 and 0.89 for PROMIS Cog Abilities and FACT Cog Abilities, respectively.

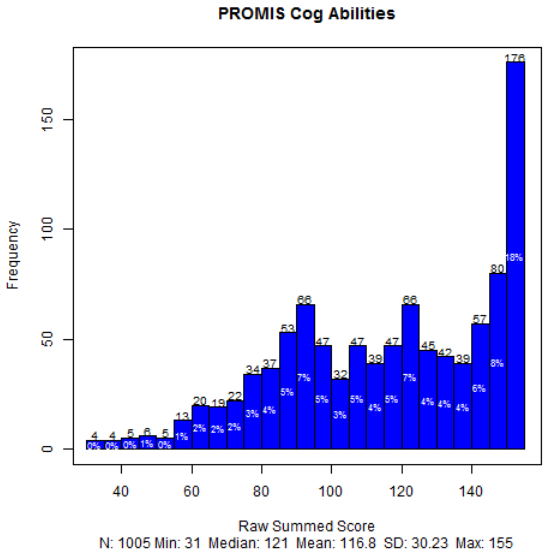


Figure 5.1.1: Raw Summed Score Distribution - PROMIS Cognitive Function - Abilities

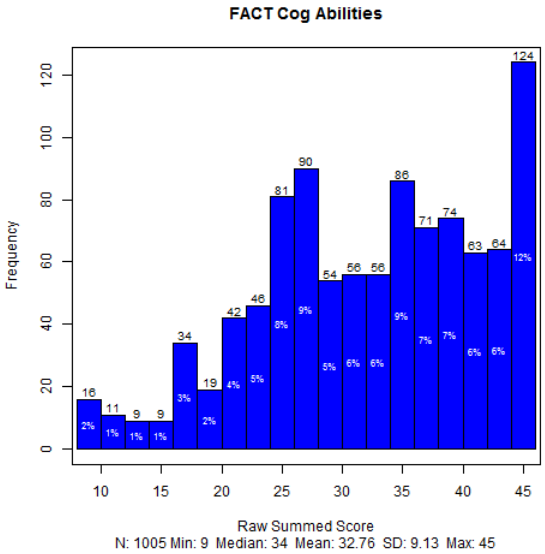


Figure 5.1.2: Raw Summed Score Distribution – FACT-Cog - Perceived Cognitive Abilities

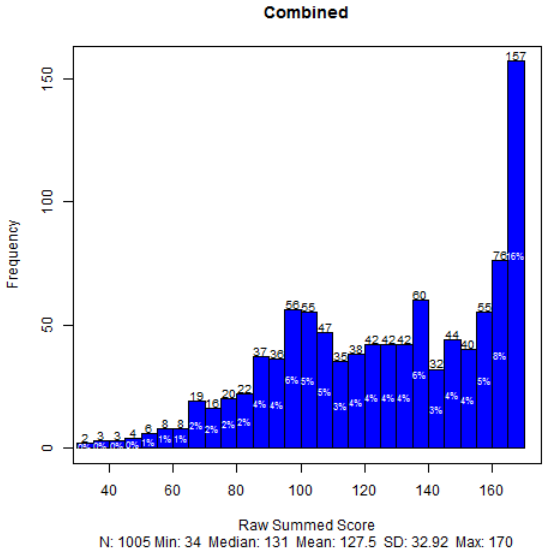


Figure 5.1.3: Raw Summed Score Distribution – Combined

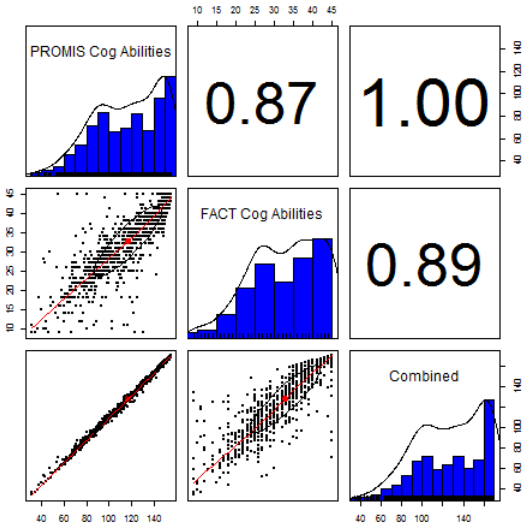


Figure 5.1.4: Scatter Plot Matrix of Raw Summed Scores

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 Cog Abilities, Cronbach's alpha internal consistency reliability estimate was 0.983 and adjusted (corrected for overlap) item-total correlations ranged from 0.652 to 0.872. For FACT Cog Abilities, alpha was 0.948 and adjusted item-total correlations ranged from 0.7 to 0.835. For the 34 items, alpha was 0.984 and adjusted item-total correlations ranged from 0.668 to 0.867.

Table 5.1.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
PROMIS Cog Abilities	31	0.983	0.652	0.802	0.872
FACT Cog Abilities	9	0.948	0.700	0.794	0.835
Combined	34	0.984	0.668	0.795	0.867

5.1.3. 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 Cog Abilities, the fit statistics were as follows: CFI = 0.958, TLI = 0.955, and RMSEA = 0.116. For FACT Cog Abilities, CFI = 0.978, TLI = 0.97, and RMSEA = 0.161. For the 34 items, CFI = 0.956, TLI = 0.953, and RMSEA =0.11. 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 Cog Abilities	31	1009	0.958	0.955	0.116
FACT Cog Abilities	9	1009	0.978	0.970	0.161
Combined	34	1009	0.956	0.953	0.110

5.1.4. 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 31 PROMIS Cog Abilities items served as anchor items to transform the item parameter estimates for the FACT Cog Abilities items onto the PROMIS Cog Abilities 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 Cog Abilities items were constrained to their final bank values, while the FACT Cog Abilities items were calibrated, under the constraints imposed by the anchor items.

Table 5.1.4: IRT Linking Constants

	A	B
Mean/Mean	1.020	-0.312
Mean/Sigma	1.100	-0.276
Haebara	1.053	-0.266
Stocking-Lord	1.080	-0.285

The item parameter estimates for the FACT Cog Abilities items were linked to the PROMIS Cog Abilities metric using the transformation constants shown in Table 5.1.5. The FACT Cog Abilities item parameter estimates from the fixed- parameter calibration are considered already on the PROMIS Cog Abilities metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for FACT Cog Abilities 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.

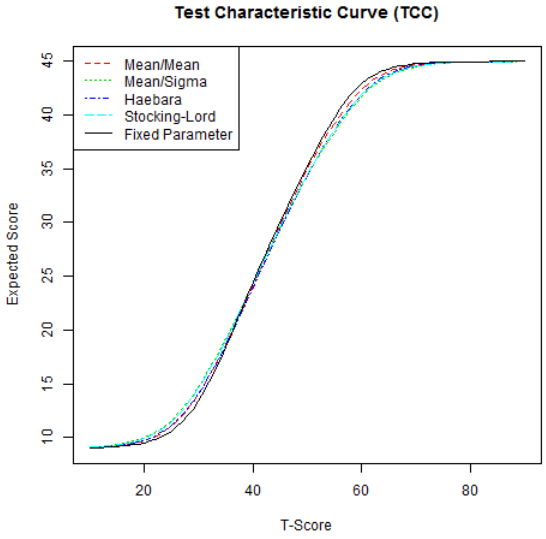


Figure 5.1.7: Test Characteristic Curves (TCC) from Different Linking Methods

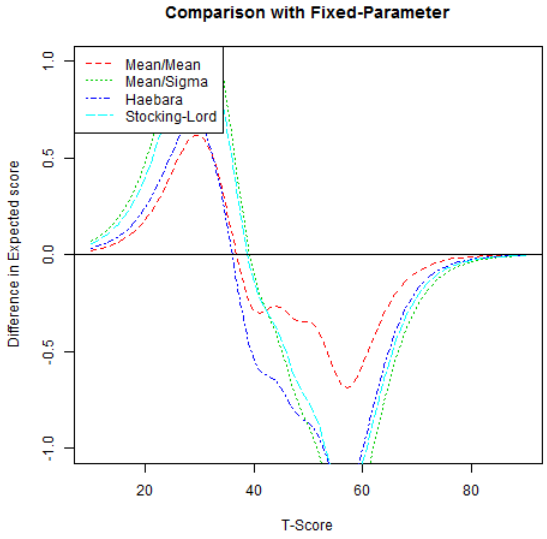


Figure 5.1.8: Difference in Test Characteristic Curves (TCC)

Table 5.1.4 shows the fixed-parameter calibration item parameter estimates for FACT Cog Abilities. The marginal reliability estimate for FACT Cog Abilities based on the item parameter estimates was 0.922. The marginal reliability estimates for PROMIS Cog Abilities and the combined set were 0.963 and 0.968, respectively. The slope parameter estimates for FACT Cog Abilities ranged from 2.26 to 4.39 with a mean of 3.15. The slope parameter estimates for PROMIS Cog Abilities ranged from 1.86 to 4.77 with a mean of 3.58. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.1.7 displays the scale information functions for PROMIS Cog Abilities, FACT Cog Abilities, and the combined set of 34. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.1.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.1.6: Fixed-Parameter Calibration Item Parameter Estimates for FACT-Cog Abilities

a	cb1	cb2	cb3	cb4	NCAT
2.490	-2.240	-1.380	-0.310	0.760	5
2.630	-2.170	-1.470	-0.710	0.170	5
4.310	-1.580	-1.110	-0.450	0.380	5
3.950	-1.640	-1.160	-0.390	0.410	5
3.630	-1.660	-1.130	-0.340	0.480	5
4.390	-1.990	-1.260	-0.420	0.350	5
2.321	-1.732	-1.012	-0.271	0.758	5
2.339	-1.723	-1.031	-0.185	0.809	5
2.265	-1.884	-1.222	-0.453	0.532	5

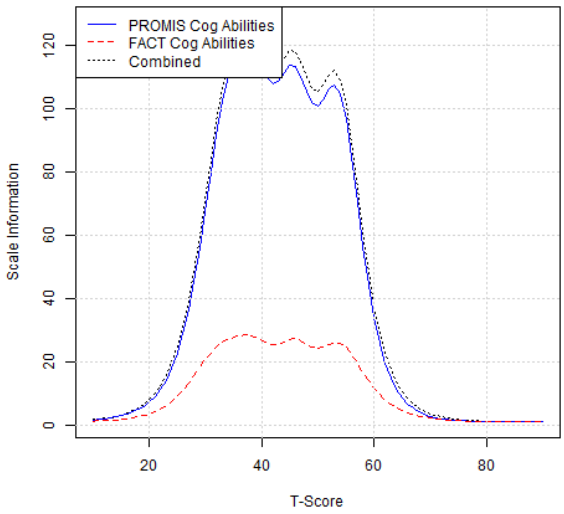


Figure 5.1.9: Comparison of Scale Information Functions

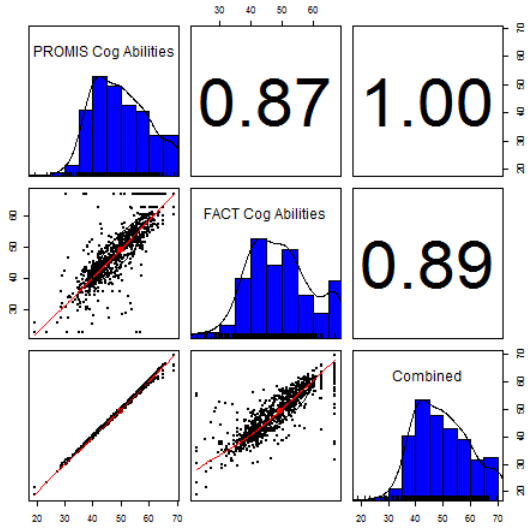


Figure 5.1.10: Comparison of IRT Scaled Scores

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

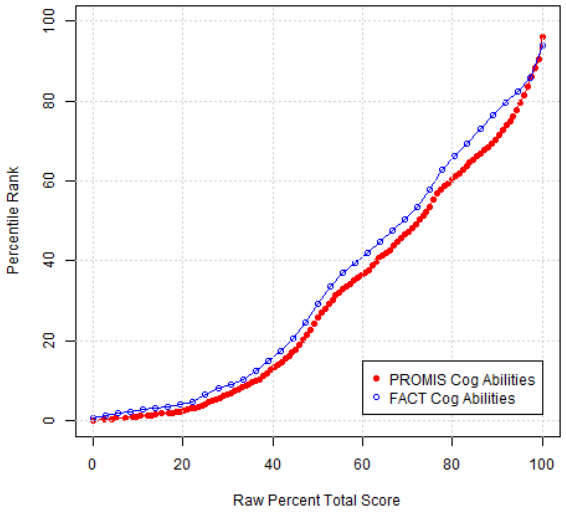


Figure 5.1.12: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

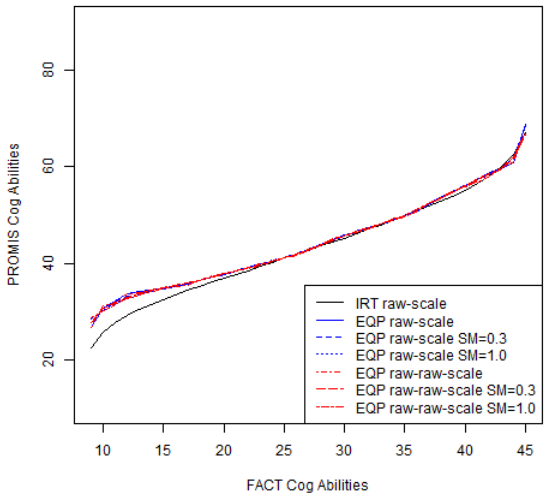


Figure 5.1.13: Equipercentile Linking Functions

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 5.1.14, 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 5.1.7 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Cog Abilities T-scores and FACT Cog Abilities scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.1.15), 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=1.0 produced the best result (0.88), followed by EQP raw-raw-scale SM=0.3 (0.879). 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 (4.804), followed by EQP raw-raw-scale SM=0.3 (4.816).

Table 5.1.8: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.796	0.452	7.387	7.397
IRT raw-scale	0.792	0.447	7.429	7.439
EQP raw-scale SM=0.0	0.785	0.251	7.555	7.556
EQP raw-scale SM=0.3	0.787	0.186	7.516	7.515
EQP raw-scale SM=1.0	0.787	0.275	7.561	7.562
EQP raw-raw-scale SM=0.0	0.789	0.126	7.408	7.406
EQP raw-raw-scale SM=0.3	0.790	0.098	7.360	7.357
EQP raw-raw-scale SM=1.0	0.797	-0.218	7.100	7.101

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1005) over a large number of replications (i.e., 10,000).

Table 5.1.9 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 Cog Abilities 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, 0.533. That is, the difference between the mean PROMIS Cog Abilities T-score and the mean equated FACT Cog Abilities T-score based on a similar sample of 75 cases is expected to be around ±1.07 (i.e., 2 × 0.533).

Table 5.1.10: 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	0.539	0.998	0.527	0.707	0.524	0.561
IRT raw-scale	0.405	1.027	0.423	0.711	0.424	0.562
EQP raw-scale SM=0.0	-0.231	0.989	-0.225	0.677	-0.217	0.553
EQP raw-scale SM=0.3	-0.264	0.981	-0.260	0.676	-0.249	0.554
EQP raw-scale SM=1.0	-0.264	0.971	-0.263	0.678	-0.258	0.549
EQP raw-raw-scale SM=0.0	-0.068	0.964	-0.083	0.675	-0.065	0.543
EQP raw-raw-scale SM=0.3	-0.010	0.956	0.006	0.663	-0.003	0.533
EQP raw-raw-scale SM=1.0	0.020	0.950	0.023	0.669	0.015	0.535

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 Cognitive Function and FACT-Cog Perceived Cognitive Impairment

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Cognition, namely the PROMIS Cognitive Function item bank (32 items) and FACT-Cog Perceived Cognitive Impairment (20 items). PROMIS Cognitive Function was scaled such that higher scores represent higher levels of Cognition. 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. Our sample consisted of 1,009 participants.

5.2.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 160 for PROMIS Cognitive Function and 100 for FACT Cog Impairment. Figures 5.2.1 and 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 Cog Function and FACT Cog Impairment was 0.89. The disattenuated (corrected for unreliabilities) correlation between PROMIS Cog Function and FACT Cog Impairment was 0.91. The correlations between the combined score and the measures were 0.98 and 0.96 for PROMIS Cog Function and FACT Cog Impairment, respectively.

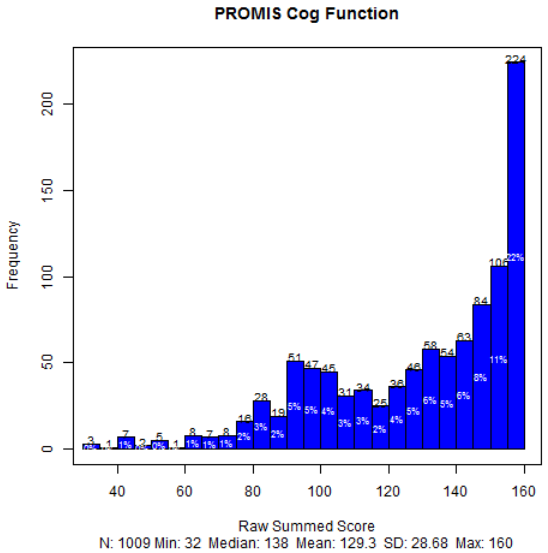


Figure 5.2.1: Raw Summed Score Distribution - PROMIS Cognitive Function

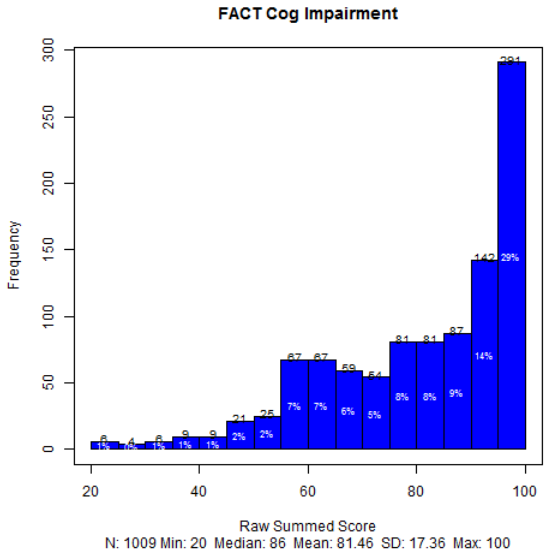


Figure 5.2.2: Raw Summed Score Distribution – FACT-Cog Perceived Cognitive Impairment

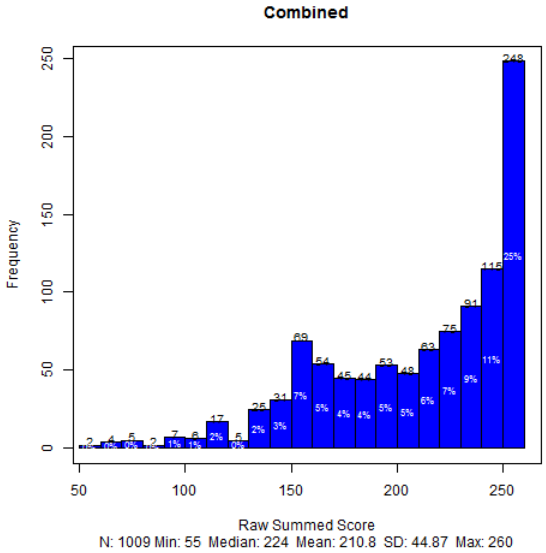


Figure 5.2.3: Raw Summed Score Distribution – Combined

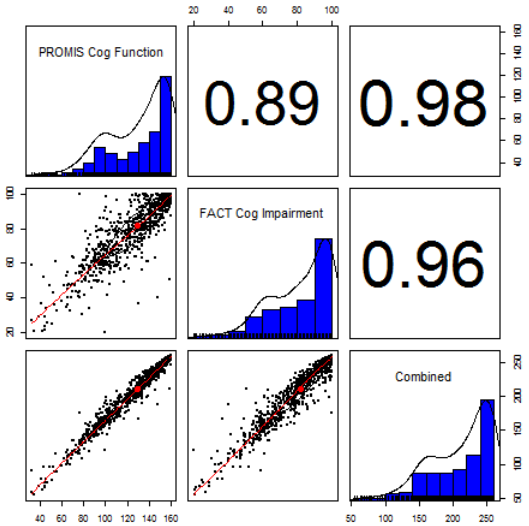


Figure 5.2.4: Scatter Plot Matrix of Raw Summed Scores

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 Cognitive Function, Cronbach's alpha internal consistency reliability estimate was 0.982 and adjusted (corrected for overlap) item-total correlations ranged from 0.643 to 0.859. For FACT Cog Impairment, alpha was 0.971 and adjusted item-total correlations ranged from 0.687 to 0.839. For the 52 items, alpha was 0.988 and adjusted item-total correlations ranged from 0.646 to 0.847.

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 Cog Function	32	0.982	0.643	0.791	0.859
FACT Cog Impairment	20	0.971	0.687	0.781	0.839
Combined	52	0.988	0.646	0.775	0.847

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 Cog Function, the fit statistics were as follows: CFI = 0.976, TLI = 0.974, and RMSEA

= 0.079. For FACT Cog Impairment, CFI = 0.976, TLI = 0.973, and RMSEA = 0.094. For the 52 items, CFI = 0.961, TLI = 0.959, and RMSEA= 0.071. 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 Cog Function	32	1009	0.976	0.974	0.079
FACT Cog Impairment	20	1009	0.976	0.973	0.094
Combined	52	1009	0.961	0.959	0.071

5.2.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 52 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 52 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 32 PROMIS Cog Function items served as anchor items to transform the item parameter estimates for the FACT-Cog Impairment items onto the PROMIS Cog 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 5.2.3 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 Cog Function items were constrained to their final bank values, while the FACT Cog Impairment items were calibrated, under the constraints imposed by the anchor items.

Table 5.2.3: IRT Linking Constants

	A	B
Mean/Mean	1.502	-0.479
Mean/Sigma	1.433	-0.510
Haebara	1.357	-0.528
Stocking-Lord	1.431	-0.500

The item parameter estimates for the FACT Cog Impairment items were linked to the PROMIS Cog Function metric using the transformation constants shown in Table 5.2.3. The FACT Cog Impairment item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Cog Function metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) FACT Cog Impairment 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.

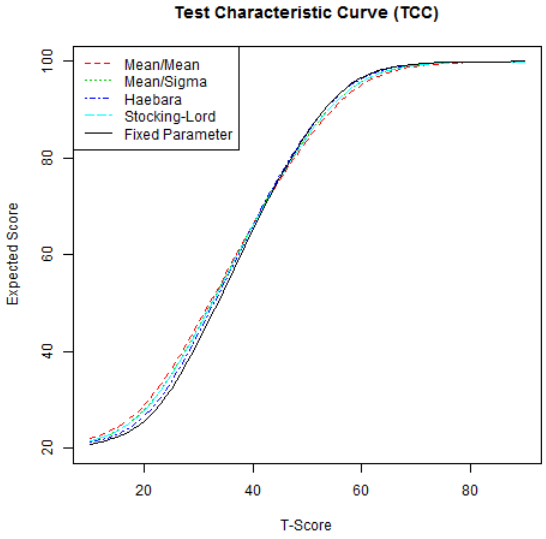


Figure 5.2.5: Test Characteristic Curves (TCC) from Different Linking Methods

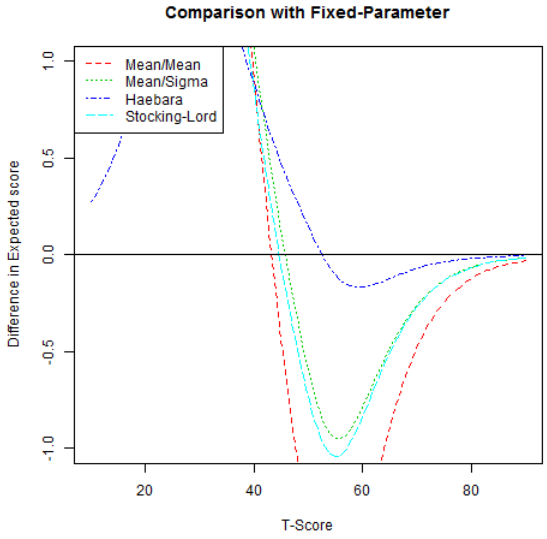


Figure 5.2.6: Difference in Test Characteristic Curves (TCC)

Table 5.2.4 Table 4 shows the fixed-parameter calibration item parameter estimates for FACT Cog Impairment. The marginal reliability estimate for FACT Cog Impairment based on the item parameter estimates was 0.937. The marginal reliability estimates for PROMIS Cog Function and the combined set were 0.956 and 0.971, respectively. The slope parameter estimates for FACT Cog Impairment ranged from 1.54 to 3.14 with a mean of 2.41. The slope parameter estimates for PROMIS Cog Function ranged from 1.34 to 3.42 with a mean of 2.36. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.2.7 displays the scale information functions for PROMIS Cog Function, FACT Cog Impairment, and the combined set of 52. 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 for FACT Cog Impairment

a	cb1	cb2	cb3	cb4	NCAT
2.230	-2.484	-1.720	-0.865	0.199	5
2.656	-2.416	-1.620	-0.839	0.187	5
2.360	-2.397	-1.457	-0.619	0.491	5
2.947	-2.270	-1.447	-0.692	0.121	5
3.052	-2.171	-1.447	-0.856	0.122	5
2.806	-2.379	-1.582	-0.753	0.060	5
1.983	-2.519	-1.517	-0.708	0.402	5
1.722	-2.784	-1.825	-0.737	0.871	5
2.703	-2.375	-1.500	-0.824	0.162	5
1.543	-2.806	-1.699	-0.783	0.625	5
3.094	-2.368	-1.601	-0.944	-0.075	5
2.297	-2.585	-1.882	-1.226	-0.573	5
1.876	-2.578	-1.790	-0.945	0.258	5
2.473	-2.314	-1.620	-0.946	0.087	5
2.579	-2.394	-1.612	-0.871	0.217	5
3.143	-2.232	-1.538	-0.896	0.036	5
2.010	-2.396	-1.560	-0.759	0.481	5
2.252	-2.580	-1.555	-0.721	0.363	5
1.956	-2.933	-2.138	-1.247	-0.213	5
2.469	-2.285	-1.595	-0.901	0.072	5

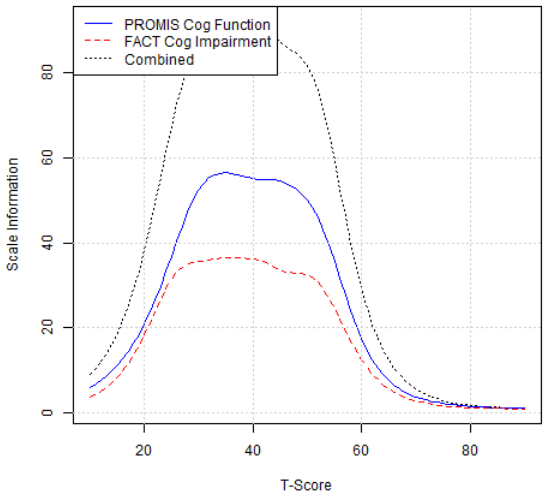


Figure 5.2.7: Comparison of Scale Information Functions

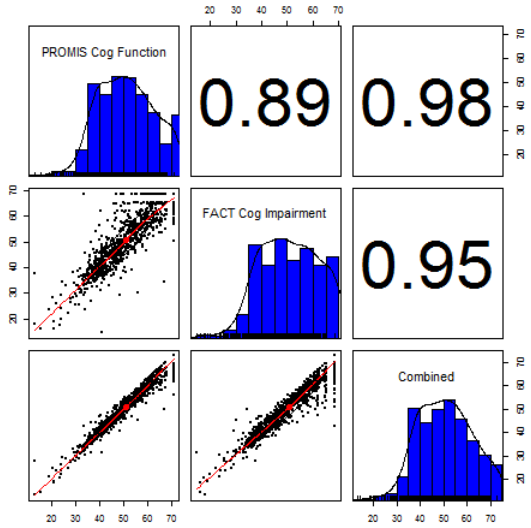


Figure 5.2.8: Comparison of IRT Scaled Scores

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 FACT Cog Impairment to a scaled score on PROMIS Cog Function can be useful. Based on the FACT- Cog Impairment 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 FACT Cog Impairment to T-score values linked to the PROMIS Cog 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.

5.2.6. Equipercentile Linking

We mapped each raw summed score point on FACT Cog Impairment to a corresponding scaled score on PROMIS Cog Function by identifying scores on PROMIS Cog Function that have the same percentile ranks as scores on FACT Cog Impairment. 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 FACT Cog Impairment to PROMIS Cog 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 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).

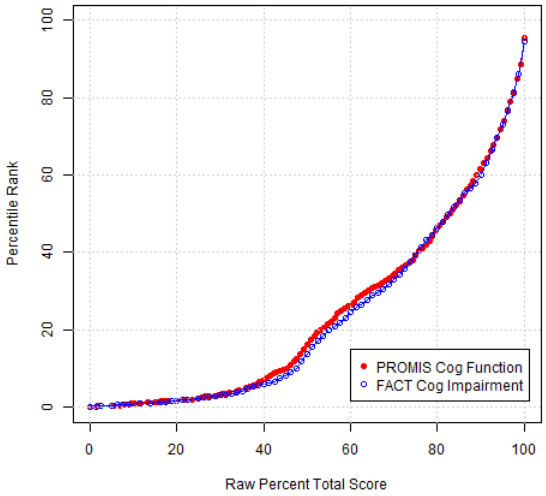


Figure 5.2.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

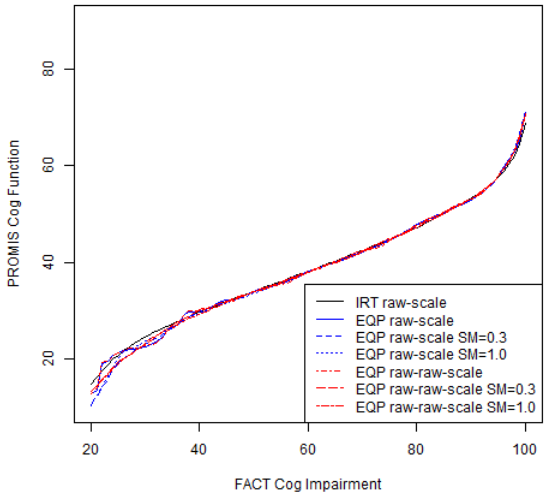


Figure 5.2.10: Equipercentile Linking Functions

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 Cog Function T-scores and FACT Cog Impairment 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 (0.886), followed by IRT raw-scale (0.882). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.475), followed by IRT raw-scale (5.547).

Table 5.2.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.886	0.187	5.475	5.475
IRT raw-scale	0.882	0.127	5.548	5.547
EQP raw-scale SM=0.0	0.880	-0.208	5.741	5.742
EQP raw-scale SM=0.3	0.879	-0.232	5.783	5.785
EQP raw-scale SM=1.0	0.879	-0.280	5.790	5.794
EQP raw-raw-scale SM=0.0	0.880	-0.200	5.721	5.721
EQP raw-raw-scale SM=0.3	0.880	-0.191	5.719	5.720
EQP raw-raw-scale SM=1.0	0.880	-0.209	5.732	5.733

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1009) 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 Cog 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, 0.619. That is, the difference between the mean PROMIS Cog Function T-score and the mean equated FACT Cog Impairment T-score based on a similar sample of 75 cases is expected to be around ±1.24 (i.e., 2 × 0.619).

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	0.179	1.075	0.190	0.762	0.187	0.619
IRT raw-scale	0.130	1.094	0.114	0.770	0.128	0.619
EQP raw-scale SM=0.0	-0.208	1.131	-0.204	0.788	-0.199	0.628
EQP raw-scale SM=0.3	-0.231	1.143	-0.246	0.808	-0.213	0.641
EQP raw-scale SM=1.0	-0.301	1.165	-0.291	0.810	-0.273	0.640
EQP raw-raw-scale SM=0.0	-0.229	1.127	-0.205	0.794	-0.188	0.630
EQP raw-raw-scale SM=0.3	-0.204	1.136	-0.191	0.787	-0.181	0.639
EQP raw-raw-scale SM=1.0	-0.210	1.123	-0.204	0.798	-0.213	0.638

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

5.3. PROMIS Cognitive Function v2.0 and Neuro-QoL Applied Cognition-General Concerns

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Cognition, namely the PROMIS Cognitive Function item bank (32 items) and Neuro-QoL Applied Cognition – General Concerns (18 items). PROMIS Cognitive Function was scaled such that higher scores represent higher levels of Cognition. 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. Our sample consisted of 1,009 participants (N = 1,008 for participants with complete responses).

Note: In 2014, the two Neuro-QoL Applied Cognition banks -- General Concerns and Executive Function – were merged into a single bank called Neuro-QoL Cognitive Function. This new bank was linked via common items to the PROMIS Cognitive Function v2.0 bank, so that the T-scores from either instrument are on the same metric. See Report 5.27 for details on the link with Neuro-QoL Cognitive Function v2.0.

5.3.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 160 for PROMIS Cog Function and 90 for Neuro-QoL Cognition. 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 Cog Function and Neuro-QoL Cognition was 0.96. The disattenuated (corrected for unreliabilities) correlation between PROMIS Cog Function and Neuro-QoL Cognition was 0.98. The correlations between the combined score and the measures were 1 and 0.98 for PROMIS Cog Function and Neuro-QoL Cognition, respectively.

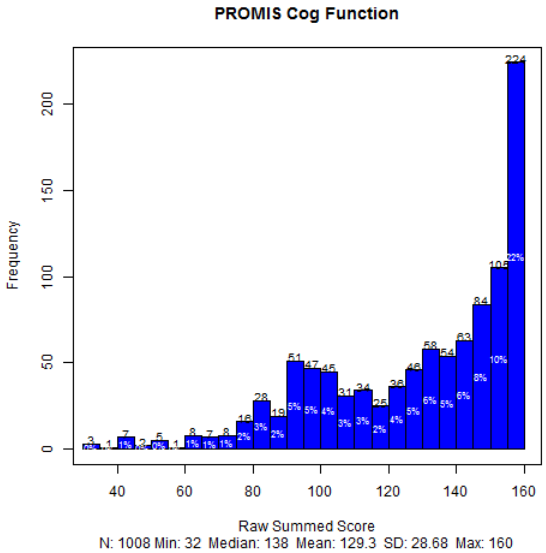


Figure 5.3.1: Raw Summed Score Distribution - PROMIS Cognitive Function

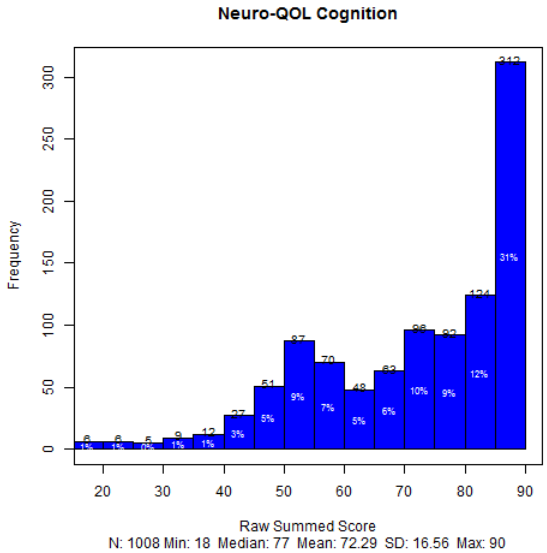


Figure 5.3.2: Raw Summed Score Distribution – Neuro-QoL Applied Cognition - General Concerns

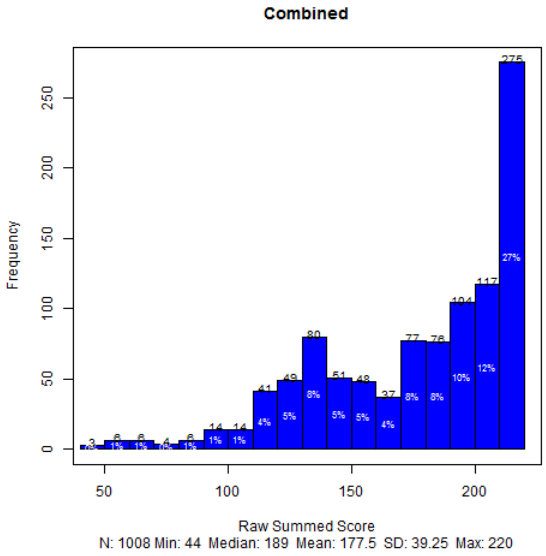


Figure 5.3.3: Raw Summed Score Distribution – Combined

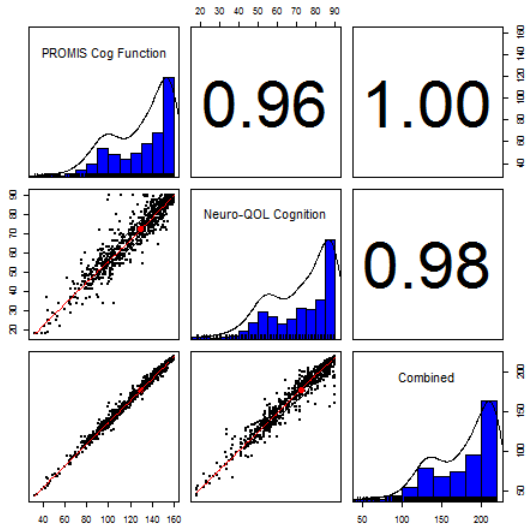


Figure 5.3.4: Scatter Plot Matrix of Raw Summed Scores

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 Cog Function, Cronbach's alpha internal consistency reliability estimate was 0.982 and adjusted (corrected for overlap) item-total correlations ranged from 0.643 to 0.859. For Neuro-QoL Cognition, alpha was 0.975 and adjusted item-total correlations ranged from 0.709 to 0.865. For the 44 items, alpha was 0.987 and adjusted item-total correlations ranged from 0.64 to 0.861.

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 Cog Function	32	0.982	0.643	0.791	0.859
Neuro-QoL Cognition	18	0.975	0.709	0.815	0.865
Combined	44	0.987	0.640	0.794	0.861

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 Cog Function, the fit statistics were as follows: CFI = 0.976, TLI = 0.974, and RMSEA = 0.079. For Neuro-QoL Cognition, CFI = 0.989, TLI = 0.987, and RMSEA = 0.08. For the 44

items, CFI = 0.975, TLI = 0.974, and RMSEA= 0.067. 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 Cog Function	32	1009	0.976	0.974	0.079
Neuro-QoL Cognition	18	1009	0.989	0.987	0.080
Combined	44	1009	0.975	0.974	0.067

5.3.4. 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 32 PROMIS Cog Function items served as anchor items to transform the item parameter estimates for the Neuro-QoL Cognition items onto the PROMIS Cog 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 5.3.3 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 Cog Function items were constrained to their final bank values, while the Neuro-QoL Cognition items were calibrated, under the constraints imposed by the anchor items.

Table 5.3.3: IRT Linking Constants

	A	B
Mean/Mean	1.554	-0.420
Mean/Sigma	1.440	-0.474
Haebara	1.369	-0.491
Stocking-Lord	1.448	-0.461

The item parameter estimates for the Neuro-QoL Cognition items were linked to the PROMIS Cog Function metric using the transformation constants shown in Table 5.3.3. The Neuro-QoL Cognition item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Cog Function metric. Neuro-QoL Cognition 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.

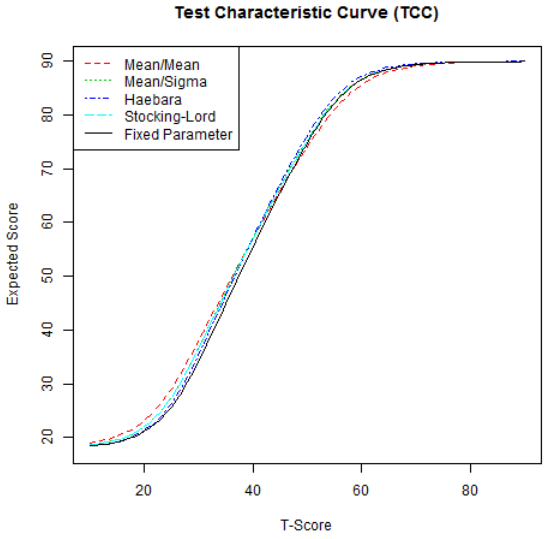


Figure 5.3.5: Test Characteristic Curves (TCC) from Different Linking Methods

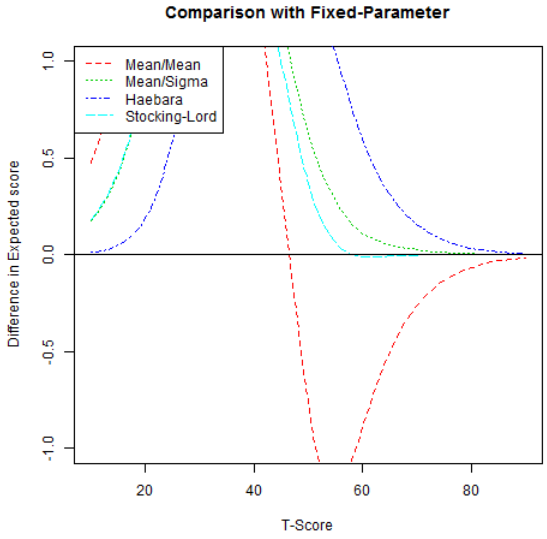


Figure 5.3.6: Difference in Test Characteristic Curves (TCC)

Table 5.3.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QOL Cognition. The marginal reliability estimate for Neuro-QoL Cognition based on the item parameter estimates was 0.941. The marginal reliability estimates for PROMIS Cog Function and the combined set were 0.956 and 0.968, respectively. The slope parameter estimates for Neuro-QoL Cognition ranged from 1.68 to 3.54 with a mean of 2.69. The slope parameter estimates for PROMIS Cog Function ranged from 1.34 to 3.42 with a mean of 2.36. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.3.7 displays the scale information functions for PROMIS Cog Function, Neuro-QoL Cognition, and the combined set of 44. 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 for Neuro-QoL Cognition

a	cb1	cb2	cb3	cb4	NCAT
2.810	-1.970	-1.580	-0.810	-0.010	5
3.090	-1.670	-1.230	-0.460	0.250	5
2.190	-2.390	-1.650	-0.560	0.350	5
2.030	-2.300	-1.710	-0.590	0.340	5
1.680	-2.420	-1.470	-0.170	1.200	5
2.870	-2.050	-1.270	-0.490	0.220	5
2.307	-2.470	-1.864	-0.890	0.390	5
1.770	-2.546	-1.639	-0.427	0.988	5
2.103	-2.555	-1.706	-0.544	0.644	5
2.991	-2.329	-1.467	-0.618	0.445	5
2.872	-2.146	-1.554	-0.609	0.276	5
2.292	-2.675	-1.753	-0.867	0.082	5
3.472	-2.103	-1.492	-0.767	0.135	5
3.544	-2.128	-1.504	-0.805	0.002	5
3.126	-2.034	-1.490	-0.594	0.296	5
3.146	-2.095	-1.490	-0.750	0.039	5
2.897	-2.125	-1.444	-0.695	0.108	5
3.299	-2.240	-1.560	-0.842	-0.050	5

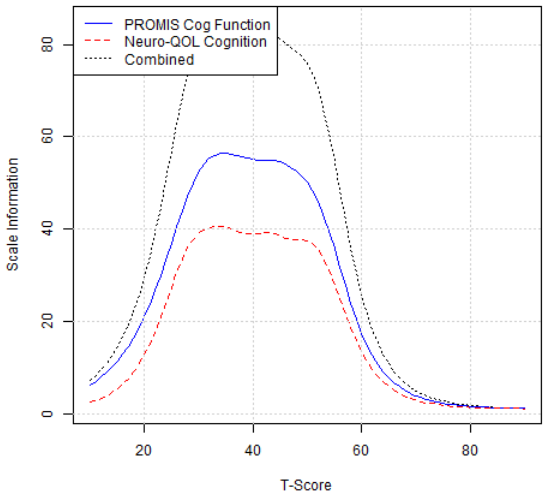


Figure 5.3.7: Comparison of Scale Information Functions

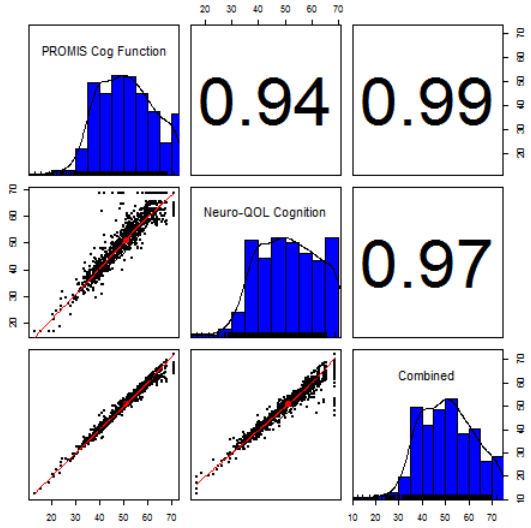


Figure 5.3.8: Comparison of IRT Scaled Scores

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 Neuro-QoL Cognition to a scaled score on PROMIS Cog Function can be useful. Based on the Neuro-QoL Cognition 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 Neuro-QoL Cognition to T-score values linked to the PROMIS Cog 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.

5.3.6. Equipercentile Linking

We mapped each raw summed score point on Neuro-QoL Cognition to a corresponding scaled score on PROMIS Cog Function by identifying scores on PROMIS Cog Function that have the same percentile ranks as scores on Neuro-QoL Cognition. 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 Neuro-QoL Cognition to PROMIS Cog 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 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).

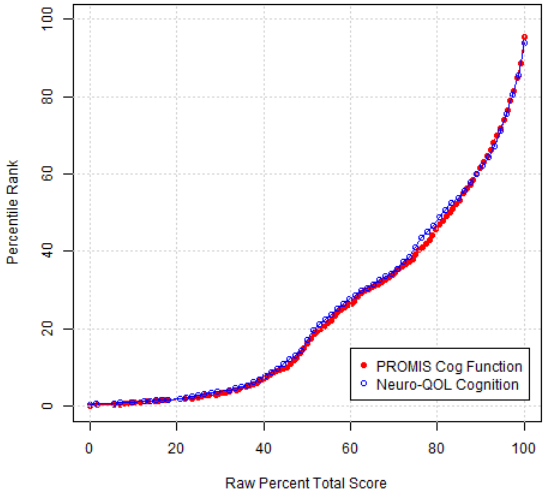


Figure 5.3.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

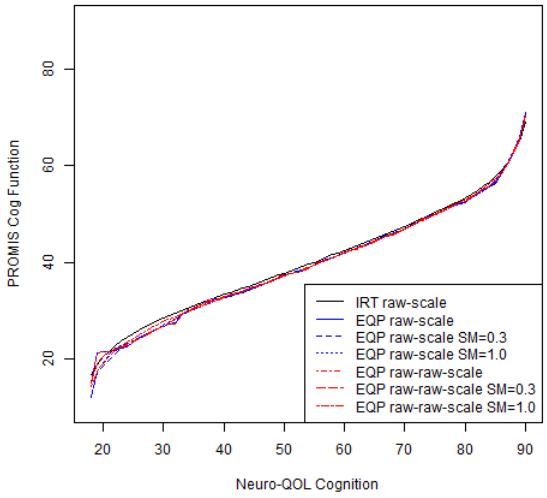


Figure 5.3.10: Equipercentile Linking Functions

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 Cog Function T-scores and Neuro-QoL Cognition 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 (0.942), followed by IRT raw-scale (0.941). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (3.94), followed by IRT raw-scale (3.972).

Table 5.3.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.942	-0.340	3.927	3.940
IRT raw-scale	0.941	-0.423	3.952	3.972
EQP raw-scale SM=0.0	0.938	-0.248	4.140	4.145
EQP raw-scale SM=0.3	0.938	-0.235	4.148	4.153
EQP raw-scale SM=1.0	0.939	-0.281	4.144	4.152
EQP raw-raw-scale SM=0.0	0.939	-0.196	4.080	4.083
EQP raw-raw-scale SM=0.3	0.939	-0.168	4.075	4.076
EQP raw-raw-scale SM=1.0	0.940	-0.169	4.066	4.067

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1008) 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 Cog 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, 0.436. That is, the difference between the mean PROMIS Cog Function T-score and the mean equated Neuro-QoL Cognition T-score based on a similar sample of 75 cases is expected to be around ±0.87 (i.e., 2 × 0.436).

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	-0.323	0.768	-0.351	0.536	-0.334	0.436
IRT raw-scale	-0.421	0.797	-0.426	0.544	-0.425	0.440
EQP raw-scale SM=0.0	-0.249	0.815	-0.242	0.562	-0.254	0.463
EQP raw-scale SM=0.3	-0.224	0.811	-0.245	0.570	-0.238	0.469
EQP raw-scale SM=1.0	-	0.813	-	0.567	-	0.466

	0.286		0.271		0.274	
EQP raw-raw-scale SM=0.0	-0.211	0.815	-0.189	0.559	-0.198	0.449
EQP raw-raw-scale SM=0.3	-0.161	0.809	-0.167	0.562	-0.168	0.458
EQP raw-raw-scale SM=1.0	-0.166	0.810	-0.161	0.555	-0.171	0.457

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 Cognitive Function and Peds PCF Short Form

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Cognition, namely the PROMIS Cog Function item bank (32 items) and Peds PCF short form (7 items). PROMIS Cog Function was scaled such that higher scores represent higher levels of Cognition. 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. Our sample consisted of 1,009 participants (N = 1,009 for participants with complete responses).

5.4.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 160 for PROMIS Cog Function and 35 for Peds PCF. 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 Cog Function and Peds PCF was 0.83. The disattenuated (corrected for unreliabilities) correlation between PROMIS Cog Function and Peds PCF was 0.87. The correlations between the combined score and the measures were 1 and 0.88 for PROMIS Cog Function and Peds PCF, respectively.

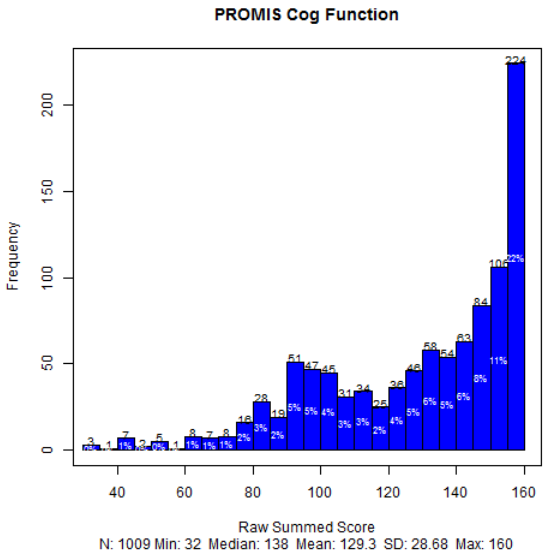


Figure 5.4.1: Raw Summed Score Distribution - PROMIS Cog Function

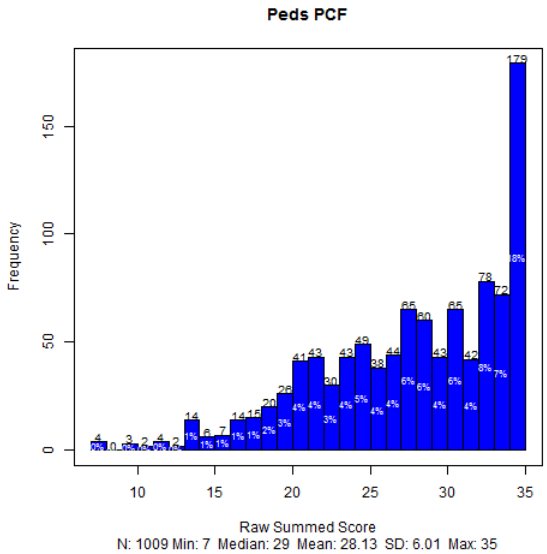


Figure 5.4.2: Raw Summed Score Distribution – Peds PCF

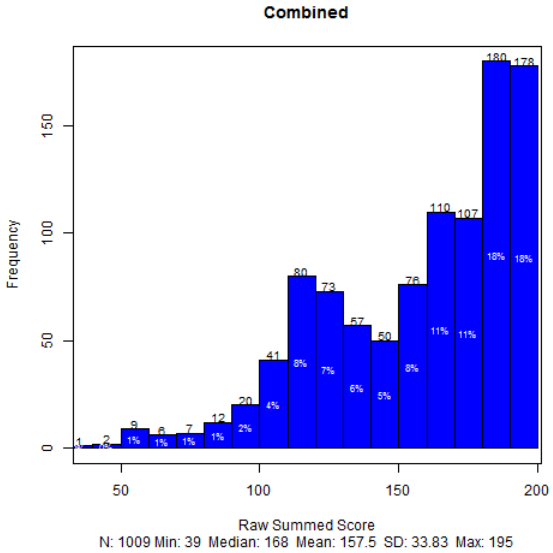


Figure 5.4.3: Raw Summed Score Distribution – Combined

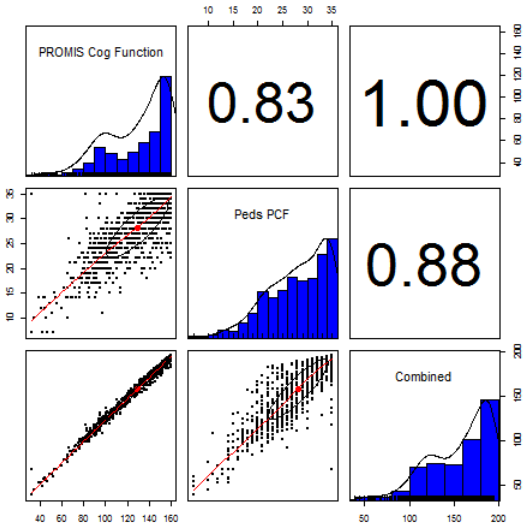


Figure 5.4.4: Scatter Plot Matrix of Raw Summed Scores

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 Cog Function, Cronbach's alpha internal consistency reliability estimate was 0.982 and adjusted (corrected for overlap) item-total correlations ranged from 0.643 to 0.859. For Peds PCF, alpha was 0.918 and adjusted item-total correlations ranged from 0.679 to 0.807. For the 39 items, alpha was 0.984 and adjusted item-total correlations ranged from 0.64 to 0.854.

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 CogFunction	32	0.982	0.643	0.791	0.859
Peds PCF	7	0.918	0.679	0.748	0.807
Combined	39	0.984	0.640	0.774	0.854

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 Cog Function, the fit statistics were as follows: CFI = 0.976, TLI = 0.974, and RMSEA = 0.079. For Peds PCF, CFI = 0.994, TLI = 0.991, and RMSEA =0.081. For the 39 items, CFI = 0.967,

TLI = 0.965, and RMSEA = 0.077. 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 Cog Function	32	1009	0.976	0.974	0.079
Peds PCF	7	1009	0.994	0.991	0.081
Combined	39	1009	0.967	0.965	0.077

5.4.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 39 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 39 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 32 PROMIS Cog Function items served as anchor items to transform the item parameter estimates for the Peds PCF items onto the PROMIS Cog 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 5.4.3 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 Cog Function items were constrained to their final bank values, while the Peds PCF items were calibrated, under the constraints imposed by the anchor items.

Table 5.4.3: IRT Linking Constants

	A	B
Mean/Mean	1.524	-0.298
Mean/Sigma	1.417	-0.358
Haebara	1.343	-0.385
Stocking-Lord	1.424	-0.344

The item parameter estimates for the Peds PCF items were linked to the PROMIS Cog Function metric using the transformation constants shown in Table 5.4.3. The Peds PCF item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Cog Function metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for Peds PCF 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.

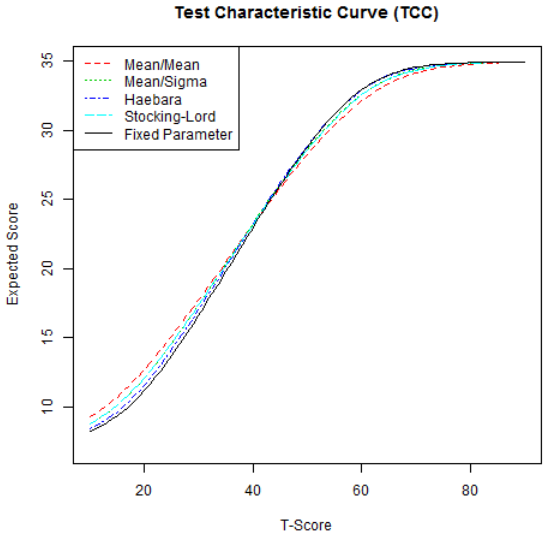


Figure 5.4.5: Test Characteristic Curves (TCC) from Different Linking Methods

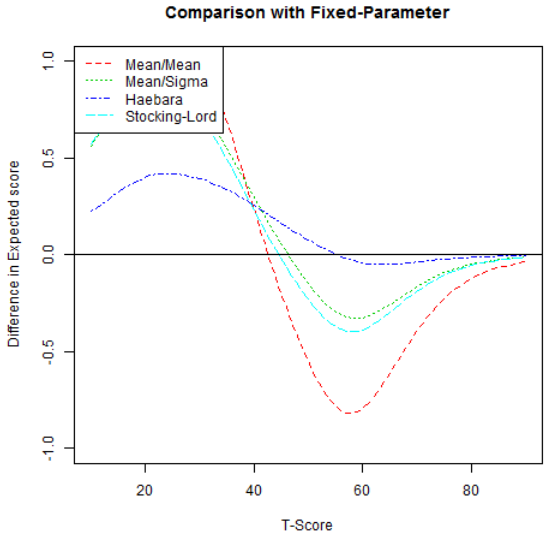


Figure 5.4.6: Difference in Test Characteristic Curves (TCC)

Table 5.4.4 shows the fixed-parameter calibration item parameter estimates for Peds PCF. The marginal reliability estimate for Peds PCF based on the item parameter estimates was 0.84. The marginal reliability estimates for PROMIS Cog Function and the combined set were 0.956 and 0.964, respectively. The slope parameter estimates for Peds PCF ranged from 1.45 to 2.39 with a mean of 1.88. The slope parameter estimates for PROMIS Cog Function ranged from 1.34 to 3.42 with a mean of 2.36. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.4.7 displays the scale information functions for PROMIS Cog Function, Peds PCF, and the combined set of 39. 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

a	cb1	cb2	cb3	cb4	NCAT
1.445	-3.604	-2.337	-0.980	0.365	5
1.837	-3.376	-1.842	-0.791	0.586	5
1.807	-2.838	-1.742	-0.770	0.590	5
1.805	-2.810	-1.856	-0.730	0.571	5
1.708	-2.608	-1.697	-0.717	0.378	5
2.157	-2.768	-1.841	-1.001	0.073	5
2.394	-2.645	-1.583	-0.796	0.244	5

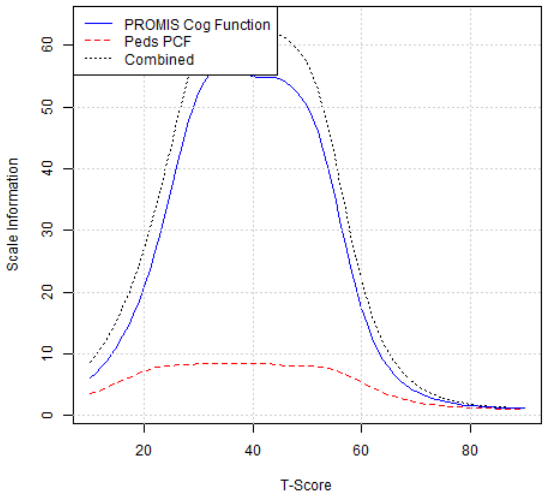


Figure 5.4.7: Comparison of Scale Information Functions

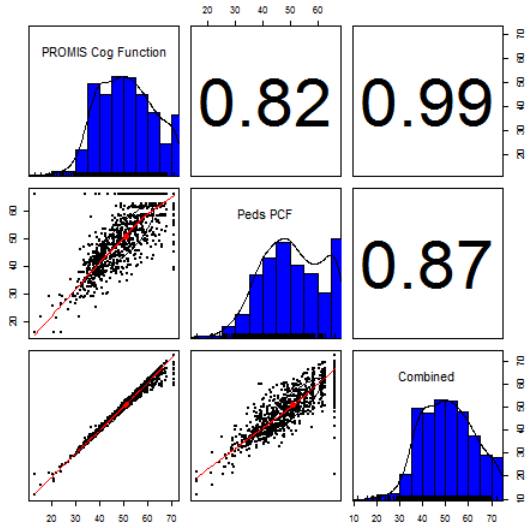


Figure 5.4.8: Comparison of IRT Scaled Scores

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 Peds PCF to a scaled score on PROMIS Cog Function can be useful. Based on the Peds PCF 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 Peds PCF to T-score values linked to the PROMIS Cog 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.

5.4.6. Equipercentile Linking

We mapped each raw summed score point on Peds PCF to a corresponding scaled score on PROMIS Cog Function by identifying scores on PROMIS Cog Function that have the same percentile ranks as scores on Peds PCF. 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 Peds PCF to PROMIS Cog 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 11and 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).

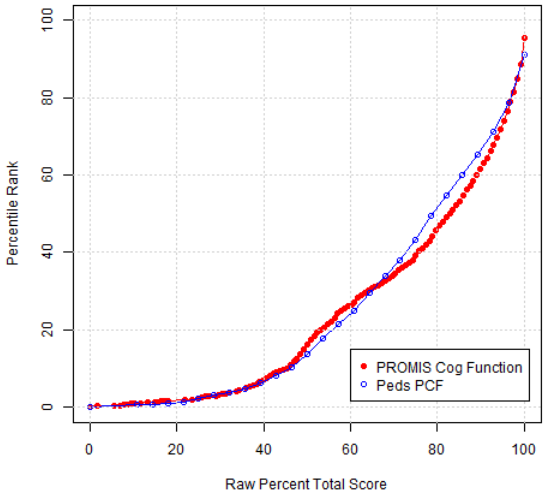


Figure 5.4.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

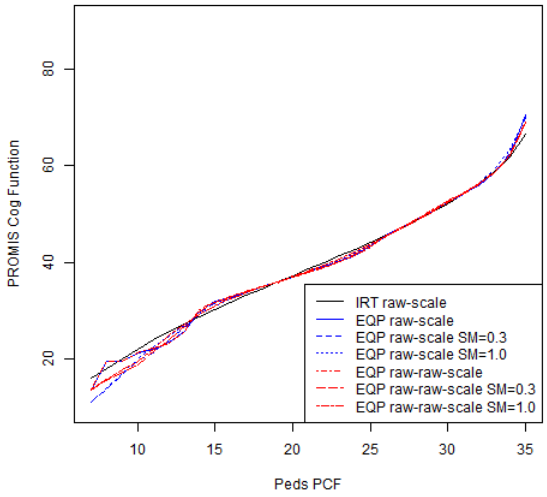


Figure 5.4.10: Equipercentile Linking Functions

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 Cog Function T-scores and Peds PCF 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 (0.815), followed by IRT raw-scale (0.812). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (6.891), followed by IRT raw-scale (6.947).

Table 5.4.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.815	0.162	6.892	6.891
IRT raw-scale	0.812	0.071	6.950	6.947
EQP raw-scale SM=0.0	0.806	-0.524	7.445	7.460
EQP raw-scale SM=0.3	0.807	-0.480	7.398	7.410
EQP raw-scale SM=1.0	0.809	-0.513	7.359	7.374
EQP raw-raw-scale SM=0.0	0.809	-0.302	7.256	7.259
EQP raw-raw-scale SM=0.3	0.809	-0.210	7.204	7.203
EQP raw-raw-scale SM=1.0	0.809	-0.259	7.229	7.230

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1009) 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 Cog 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, 0.778. That is, the difference between the mean PROMIS Cog Function T-score and the mean equated Peds PCF T-score based on a similar sample of 75 cases is expected to be around ±1.56 (i.e., 2 × 0.778)

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	0.181	1.362	0.180	0.951	0.160	0.779
IRT raw-scale	0.043	1.372	0.063	0.963	0.070	0.778
EQP raw-scale SM=0.0	-0.529	1.472	-0.536	1.024	-0.526	0.826
EQP raw-scale SM=0.3	-0.470	1.463	-0.462	1.017	-0.484	0.822
EQP raw-scale SM=1.0	-0.504	1.458	-0.501	1.007	-0.504	0.824
EQP raw-raw-scale SM=0.0	-0.275	1.444	-0.312	1.006	-0.299	0.800
EQP raw-raw-scale SM=0.3	-0.229	1.422	-0.224	0.978	-0.194	0.805
EQP raw-raw-scale SM=1.0	-0.304	1.422	-0.264	1.001	-0.257	0.790

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 Anxiety and HADS

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 (a selection of 15 highly informative items) and HADS Anxiety (7 items). PROMIS Anxiety and HADS Anxiety were 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. Our sample consisted of 1,120 participants (N = 1,015 for participants with complete responses).

5.5.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 75 for PROMIS Anxiety and 28 for HADS Anxiety. 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 Anxiety and HADS Anxiety was 0.67. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anxiety and HADS Anxiety was 0.74. The correlations between the combined score and the measures were 0.98 and 0.81 for PROMIS Anxiety and HADS Anxiety, respectively.

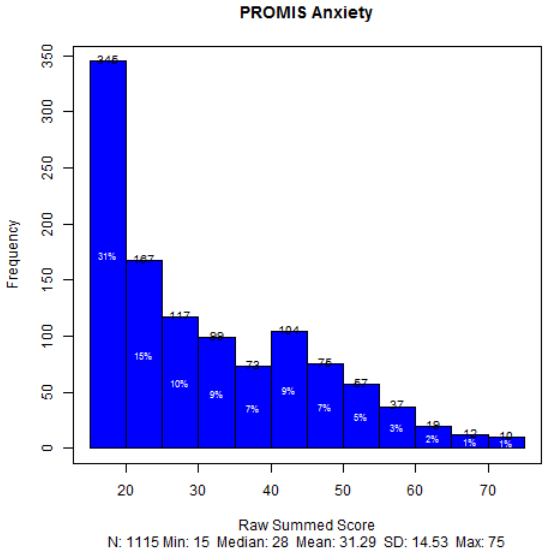


Figure 5.5.1: Raw Summed Score Distribution - PROMIS Anxiety

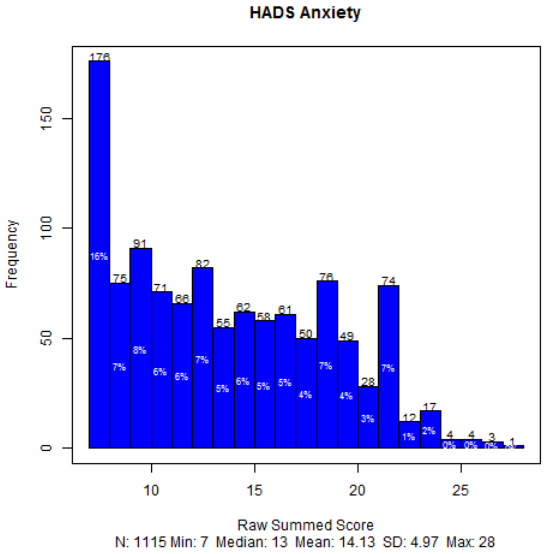


Figure 5.5.2: Raw Summed Score Distribution – HADS Anxiety

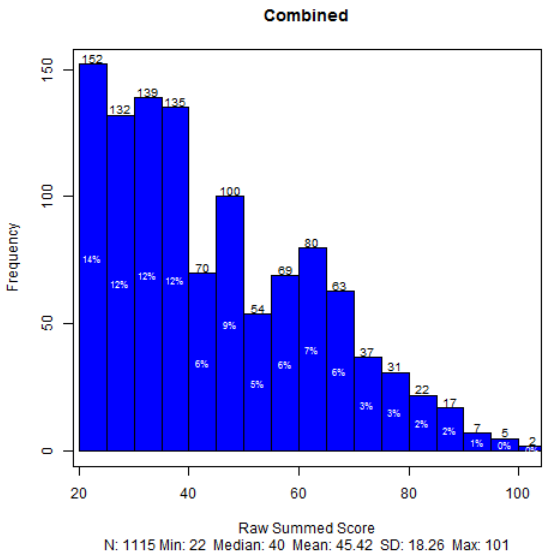


Figure 5.5.3: Raw Summed Score Distribution – Combined

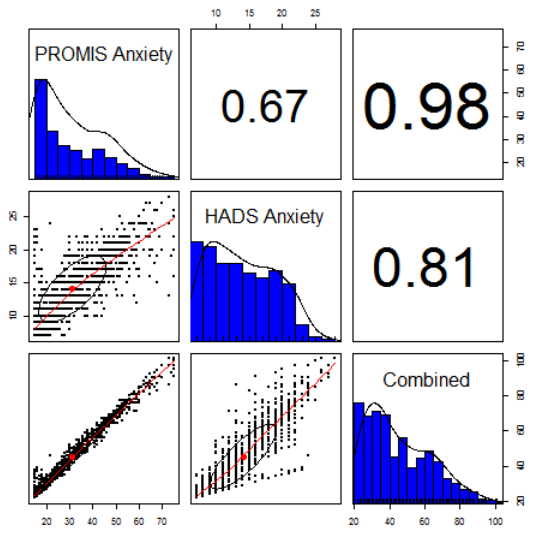


Figure 5.5.4: Scatter Plot Matrix of Raw Summed Scores

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 Anxiety, Cronbach's alpha internal consistency reliability estimate was 0.975 and adjusted (corrected for overlap) item-total correlations ranged from 0.792 to 0.88. For HADS Anxiety, alpha was 0.862 and adjusted item-total correlations ranged from 0.398 to 0.772. For the 22 items, alpha was 0.968 and adjusted item-total correlations ranged from 0.452 to 0.863.

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 Anxiety	15	0.975	0.792	0.838	0.880
HADS Anxiety	7	0.862	0.398	0.630	0.772
Combined	22	0.968	0.452	0.743	0.863

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 Anxiety, the fit statistics were as follows: CFI = 0.985, TLI = 0.982, and RMSEA = 0.116. For HADS Anxiety, CFI = 0.943, TLI = 0.915, and RMSEA= 0.21. For the 22 items, CFI = 0.936, TLI = 0.929, and RMSEA = 0.168. 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 Anxiety	15	1120	0.985	0.982	0.116
HADS Anxiety	7	1120	0.943	0.915	0.210
Combined	22	1120	0.936	0.929	0.168

5.5.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 22 items according to the graded response model. The calibration was run using `MULTILOG` and two different approaches as described previously (i.e., IRT linking vs. fixed-parameter calibration). For IRT linking, all 22 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 15 PROMIS Anxiety items served as anchor items to transform the item parameter estimates for the HADS 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 5.5.3 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 Anxiety items were constrained to their final bank values, while the HADS Anxiety items were calibrated, under the constraints imposed by the anchor items.

Table 5.5.3: IRT Linking Constants

	A	B
Mean/Mean	1.190	0.410
Mean/Sigma	1.253	0.364
Haebara	1.235	0.389
Stocking-Lord	1.244	0.367

The item parameter estimates for the HADS Anxiety items were linked to the PROMIS Anxiety metric using the transformation constants shown in Table 5.5.3. The HADS 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 HADS Anxiety 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.

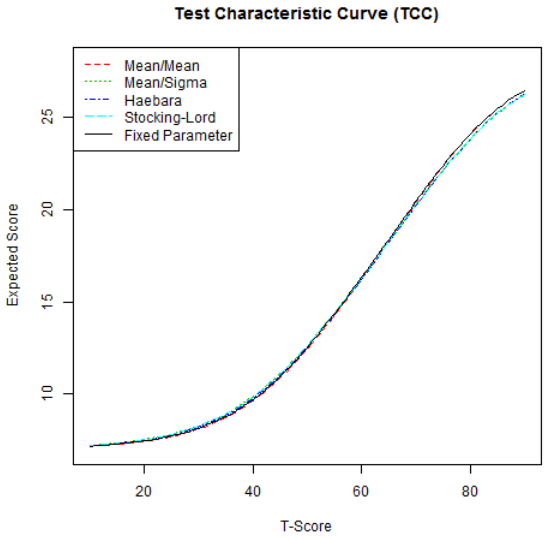


Figure 5.5.5: Test Characteristic Curves (TCC) from Different Linking Methods

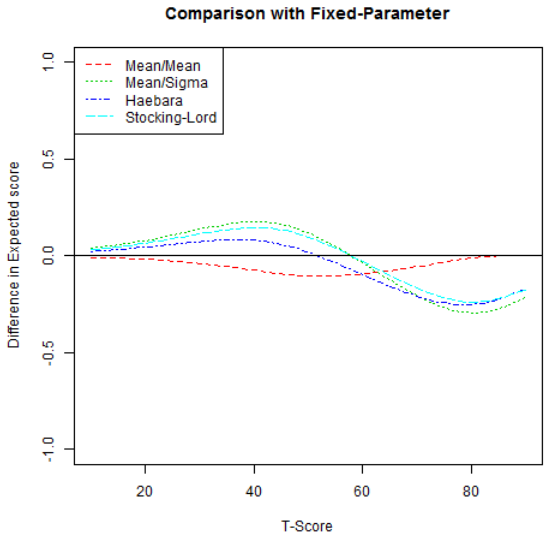


Figure 5.5.6: Difference in Test Characteristic Curves (TCC)

Table 5.5.4 shows the fixed-parameter calibration item parameter estimates for HADS Anxiety. The marginal reliability estimate for HADS Anxiety based on the item parameter estimates was 0.733. The marginal reliability estimates for PROMIS Anxiety and the combined set were 0.922 and 0.94, respectively. The slope parameter estimates for HADS Anxiety ranged from 0.863 to 1.93 with a mean of 1.29. The slope parameter estimates for PROMIS Anxiety ranged from 2.41 to 3.88 with a mean of 3.24. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.5.7 displays the scale information functions for PROMIS Anxiety, HADS Anxiety, and the combined set of 22. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.5.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.5.4: Fixed-Parameter Calibration Item Parameter Estimates

a	cb1	cb2	cb3	NCAT
0.954	-1.088	1.352	2.674	4
1.156	-0.121	1.150	2.451	4
1.178	-0.704	0.973	2.293	4
1.604	-0.483	1.116	2.787	4
1.933	0.252	1.792	2.997	4
0.863	-0.629	1.491	3.257	4
1.349	0.289	1.547	2.660	4

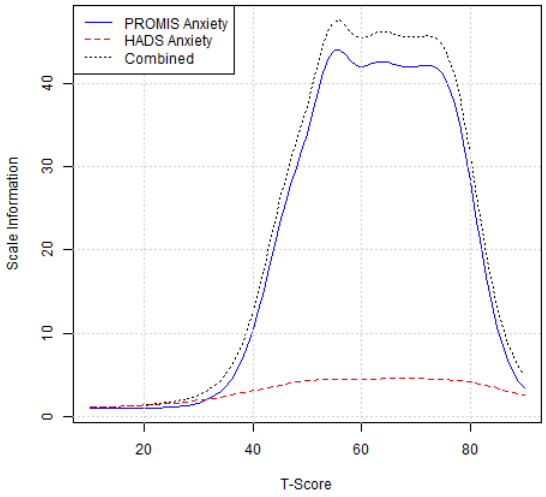


Figure 5.5.7: Comparison of Scale Information Functions

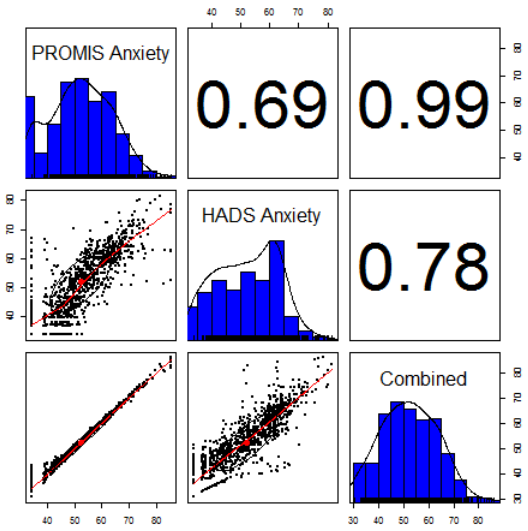


Figure 5.5.8: Comparison of IRT Scaled Scores

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 HADS Anxiety to a scaled score on PROMIS Anxiety can be useful. Based on the HADS Anxiety 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 HADS 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.

5.5.6. Equipercentile Linking

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

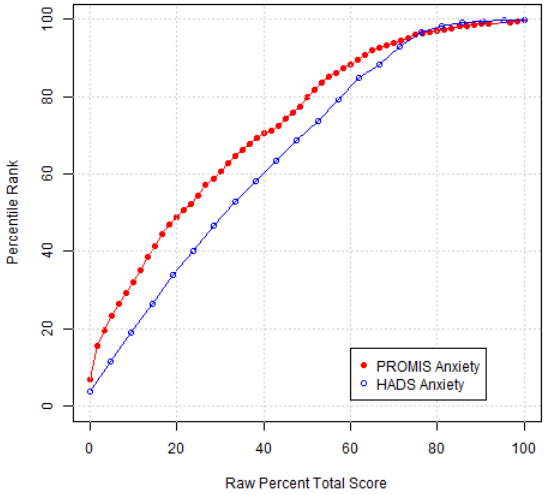


Figure 5.5.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

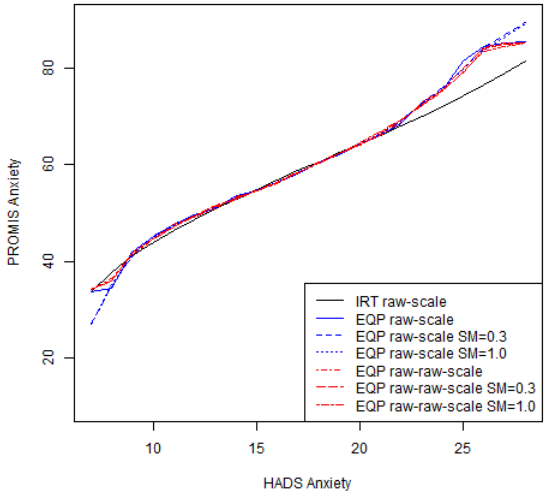


Figure 5.5.10: Equipercentile Linking Functions

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 Anxiety T-scores and HADS Anxiety 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 (0.689), followed by IRT raw-scale (0.621). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (8.728), followed by IRT raw-scale (9.695).

Table 5.5.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.689	0.177	8.730	8.728
IRT raw-scale	0.621	0.261	9.696	9.695
EQP raw-scale SM=0.0	0.618	0.134	9.972	9.969
EQP raw-scale SM=0.3	0.617	0.576	10.440	10.451
EQP raw-scale SM=1.0	0.617	0.578	10.449	10.460
EQP raw-raw-scale SM=0.0	0.617	0.032	9.891	9.887
EQP raw-raw-scale SM=0.3	0.614	0.028	9.956	9.951
EQP raw-raw-scale SM=1.0	0.613	0.029	9.986	9.982

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1115) 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 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, 0.971. That is, the difference between the mean PROMIS Anxiety T-score and the mean equated HADS Anxiety T-score based on a similar sample of 75 cases is expected to be around ±1.94 (i.e., 2 × 0.971).

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	0.179	1.728	0.174	1.210	0.197	0.971
IRT raw-scale	0.239	1.921	0.231	1.335	0.258	1.078
EQP raw-scale SM=0.0	0.170	1.978	0.147	1.376	0.145	1.112
EQP raw-scale SM=0.3	0.568	2.080	0.560	1.446	0.600	1.155
EQP raw-scale SM=1.0	0.584	2.058	0.574	1.435	0.590	1.157
EQP raw-raw-scale SM=0.0	0.001	1.983	0.014	1.365	0.033	1.114
EQP raw-raw-scale SM=0.3	0.026	1.978	0.023	1.383	0.024	1.108
EQP raw-raw-scale SM=1.0	0.050	1.970	0.047	1.380	0.052	1.103

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 Anxiety and PANAS-Negative Affect

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 (a selection of 15 highly informative items) and PANAS Negative Affect (10 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. Our sample consisted of 1,120 participants (N = 1,109 for participants with complete responses).

5.6.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 75 for PROMIS Anxiety and 50 for PANAS NA. 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 Anxiety and PANAS NA was 0.89. The disattenuated (corrected for unreliabilities) correlation between PROMIS Anxiety and PANAS NA was 0.93. The correlations between the combined score and the measures were 0.98 and 0.96 for PROMIS Anxiety and PANAS NA, respectively.

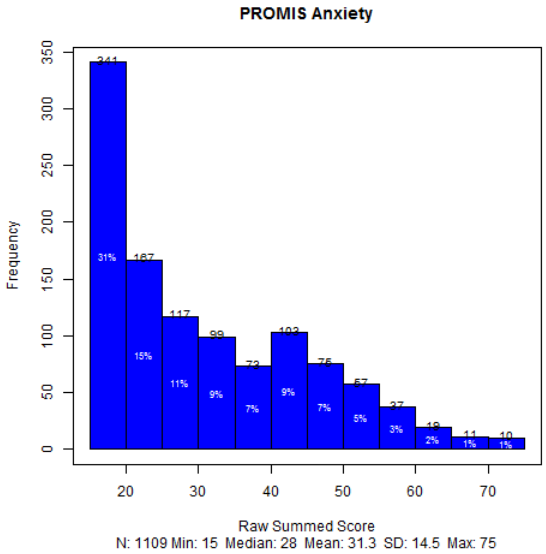


Figure 5.6.1: Raw Summed Score Distribution - PROMIS Anxiety

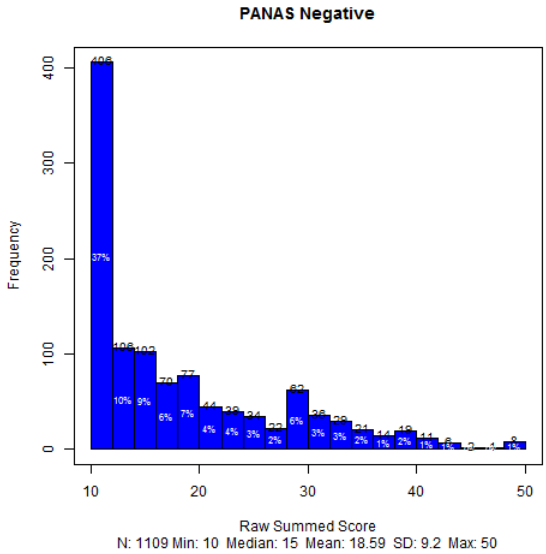


Figure 5.6.2: Raw Summed Score Distribution – PANAS NA

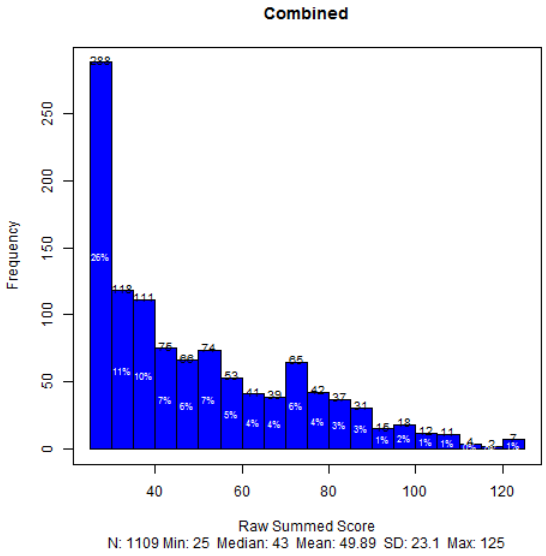


Figure 5.6.3: Raw Summed Score Distribution – Combined

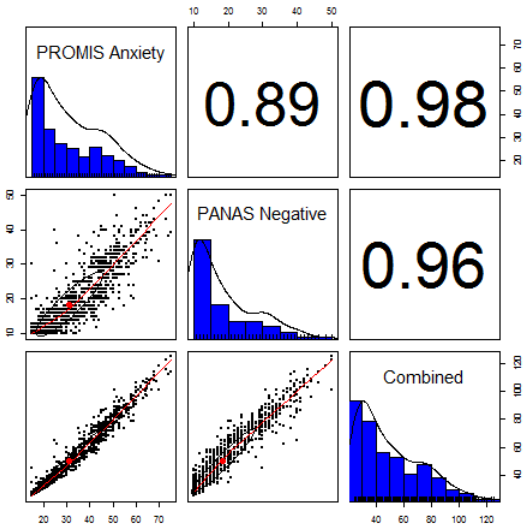


Figure 5.6.4: Scatter Plot Matrix of Raw Summed Scores

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 Anxiety, Cronbach's alpha internal consistency reliability estimate was 0.975 and adjusted (corrected for overlap) item-total correlations ranged from 0.792 to 0.88. For PANAS NA, alpha was 0.954 and adjusted item-total correlations ranged from 0.74 to 0.844. For the 25 items, alpha was 0.981 and adjusted item-total correlations ranged from 0.69 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 Anxiety	15	0.975	0.792	0.838	0.880
PANAS NA	10	0.954	0.740	0.801	0.844
Combined	25	0.981	0.690	0.815	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 Anxiety, the fit statistics were as follows: CFI = 0.985, TLI = 0.982, and RMSEA = 0.116. For PANAS NA, CFI = 0.984, TLI = 0.98, and RMSEA= 0.125. For the 25 items, CFI = 0.975, TLI = 0.972, and RMSEA = 0.102. 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 Anxiety	15	1120	0.985	0.982	0.116
PANAS NA	10	1120	0.984	0.980	0.125
Combined	25	1120	0.975	0.972	0.102

5.6.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 25 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 25 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 PANAS NA 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 5.6.3 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 Anxiety items were constrained to their final bank values, while the PANAS NA items were calibrated, under the constraints imposed by the anchor items.

Table 5.6.3: IRT Linking Constants

	A	B
Mean/Mean	1.254	0.568
Mean/Sigma	1.348	0.515
Haebara	1.326	0.540
Stocking-Lord	1.334	0.519

The item parameter estimates for the PANAS NA items were linked to the PROMIS Anxiety metric using the transformation constants shown in Table 5.6.3. The PANAS NA 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 PANAS NA 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.

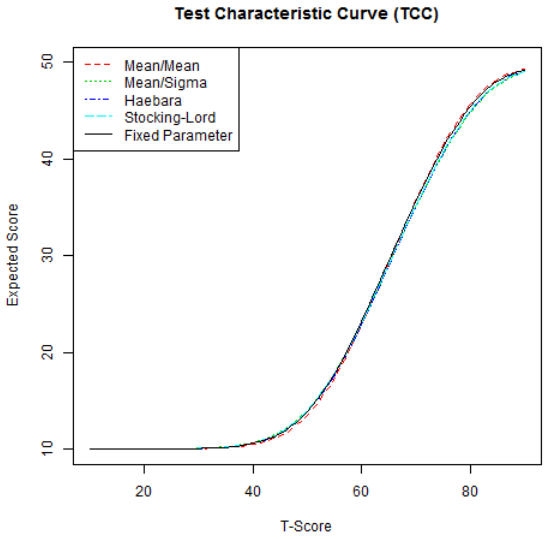


Figure 5.6.5: Test Characteristic Curves (TCC) from Different Linking Methods

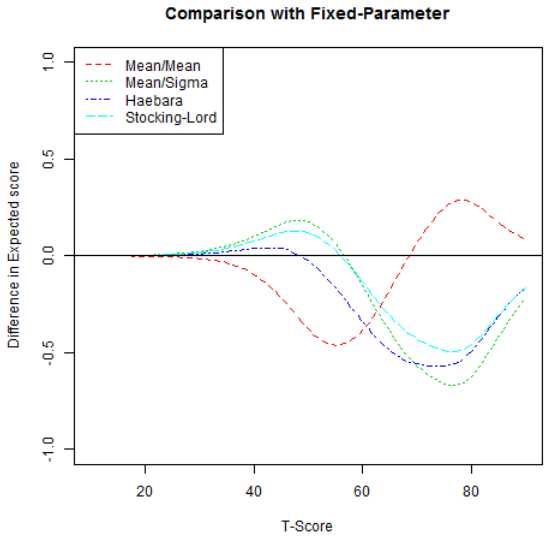


Figure 5.6.6: Difference in Test Characteristic Curves (TCC)

Table 5.6.4 shows the fixed-parameter calibration item parameter estimates for PANAS NA. The marginal reliability estimate for PANAS NA based on the item parameter estimates was 0.843. The marginal reliability estimates for PROMIS Anxiety and the combined set were 0.922 and 0.937, respectively. The slope parameter estimates for PANAS NA ranged from 1.82 to 3.4 with a mean of 2.6. The slope parameter estimates for PROMIS Anxiety ranged from 2.41 to 3.88 with a mean of 3.24. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.6.7 displays the scale information functions for PROMIS Anxiety, PANAS NA, and the combined set of 25. 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

a	cb1	cb2	cb3	cb4	NCAT
2.693	0.032	0.996	1.798	2.852	5
2.392	-0.158	0.954	1.780	2.986	5
2.451	0.534	1.231	2.006	2.916	5
2.860	0.598	1.329	1.967	2.761	5
1.821	0.617	1.564	2.380	3.501	5
2.012	-0.326	0.965	1.792	2.783	5
2.479	0.707	1.353	1.970	2.718	5
3.218	0.136	0.989	1.601	2.387	5
2.690	0.421	1.229	1.875	2.677	5
3.403	0.618	1.275	1.805	2.568	5

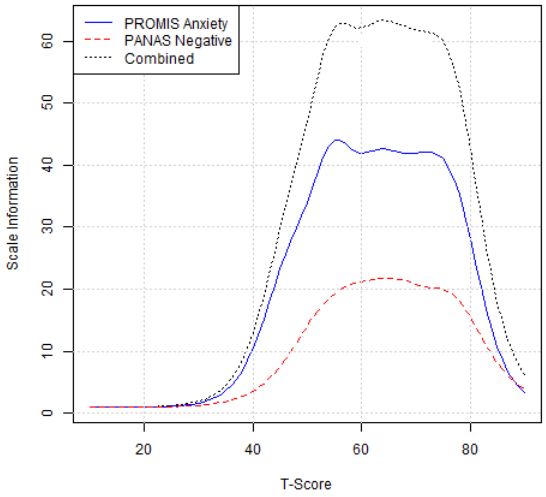


Figure 5.6.7: Comparison of Scale Information Functions

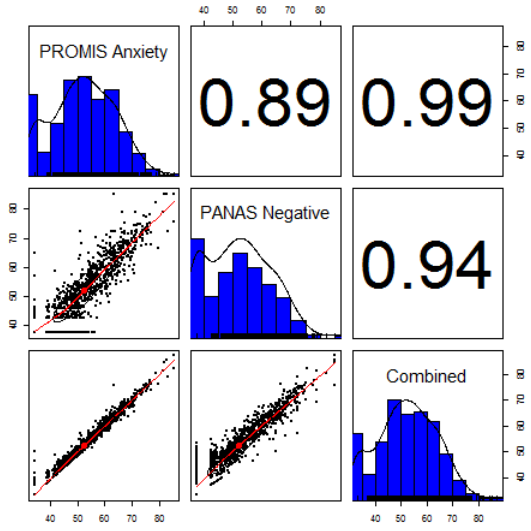


Figure 5.6.8: Comparison of IRT Scaled Scores

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 PANAS NA to a scaled score on PROMIS Anxiety can be useful. Based on the PANAS NA 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 PANAS NA 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.

5.6.6. Equipercentile Linking

We mapped each raw summed score point on PANAS NA to a correspond- ing scaled score on PROMIS Anxiety by identifying scores on PROMIS Anxiety that have the same percentile ranks as scores on PANAS NA. 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 5.6.9 displays the cumulative distribution functions of the measures. Figure 5.6.10 shows the equipercentile linking functions based on raw summed scores, from PANAS NA

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 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).

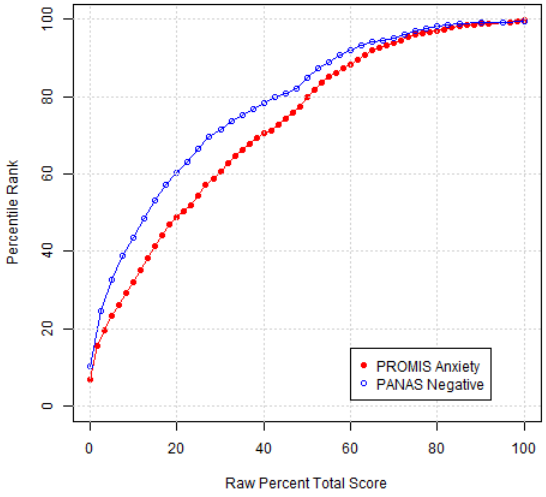


Figure 5.6.10: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

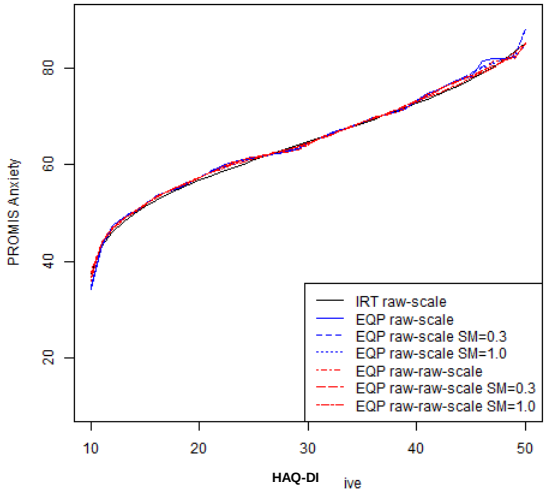


Figure 5.6.11: Equipercentile Linking Functions

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, With respect to the correlation between observed and linked T-scores, IRT pattern scoring produced the best result (0.894), followed by EQP raw-raw-scale SM=1.0 (0.89). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.165), followed by EQP raw-raw- scale SM=1.0 (5.24).

Table 5.6.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Anxiety T-scores and PANAS NA 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 (0.894), followed by EQP raw-raw-scale SM=1.0 (0.89). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.165), followed by EQP raw-raw- scale SM=1.0 (5.24).

Table 5.6.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.894	0.102	5.167	5.165
IRT raw-scale	0.890	0.255	5.256	5.260
EQP raw-scale SM=0.0	0.885	0.480	5.610	5.628
EQP raw-scale SM=0.3	0.885	0.417	5.578	5.591
EQP raw-scale SM=1.0	0.887	0.381	5.514	5.525
EQP raw-raw-scale SM=0.0	0.888	0.163	5.385	5.385
EQP raw-raw-scale SM=0.3	0.889	0.036	5.312	5.310
EQP raw-raw-scale SM=1.0	0.890	-0.136	5.240	5.240

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1009) 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 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, 0.579. That is, the difference between the mean PROMIS Anxiety T-score and the mean equated PANAS NA T-score based on a similar sample of 75 cases is expected to be around ±1.16 (i.e., 2 × 0.579).

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	0.114	1.021	0.099	0.712	0.103	0.582
IRT raw-scale	0.250	1.053	0.259	0.718	0.252	0.586
EQP raw-scale SM=0.0	0.481	1.111	0.489	0.781	0.469	0.623
EQP raw-scale SM=0.3	0.404	1.112	0.409	0.779	0.426	0.624
EQP raw-scale SM=1.0	0.384	1.101	0.377	0.756	0.374	0.621
EQP raw-raw-scale SM=0.0	0.169	1.065	0.160	0.747	0.168	0.606
EQP raw-raw-scale SM=0.3	0.035	1.049	0.029	0.720	0.041	0.596
EQP raw-raw-scale SM=1.0	-0.140	1.034	-0.134	0.728	-0.133	0.579

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 Depression and BDI-II

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression item bank (a selection of 15 highly informative items) and BDI-II (21 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. Our sample consisted of 1,120 participants (N = 1,104 for participants with complete responses).

5.7.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 75 for PROMIS Depression and 84 for BDI-II. 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 Depression and BDI-II was 0.89. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and BDI-II was 0.91. The correlations between the combined score and the measures were 0.97 and 0.97 for PROMIS Depression and BDI-II, respectively.

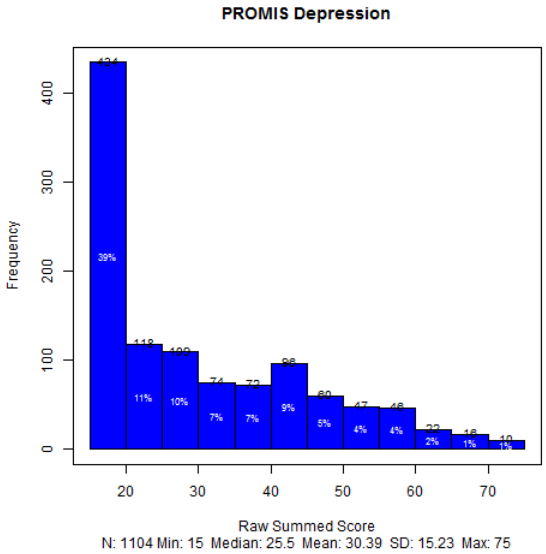


Figure 5.7.1: Raw Summed Score Distribution - PROMIS Depression

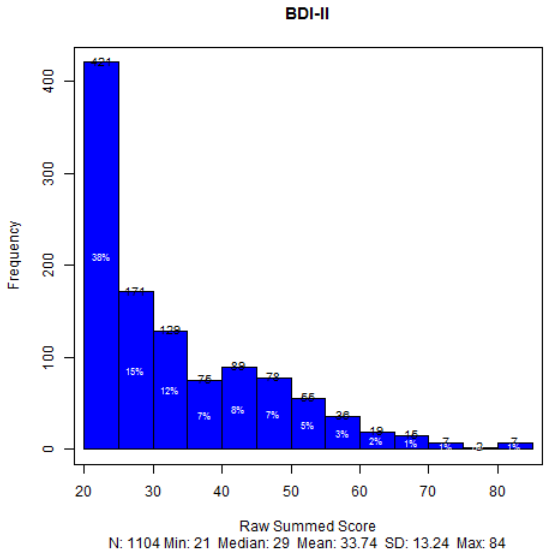


Figure 5.7.2: Raw Summed Score Distribution – BDI-II

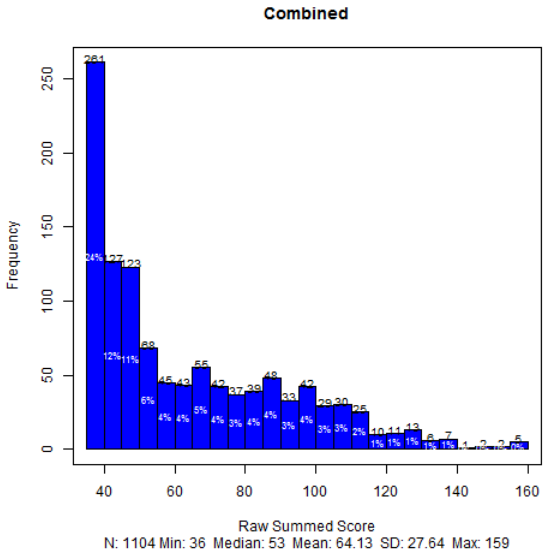


Figure 5.7.3: Raw Summed Score Distribution – Combined

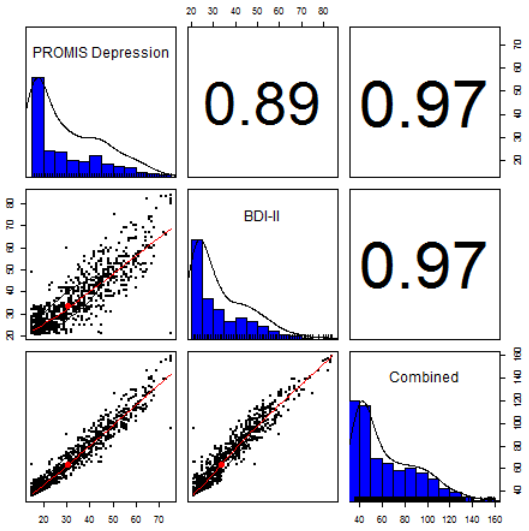


Figure 5.7.4: Scatter Plot Matrix of Raw Summed Scores

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 Depression, Cronbach's alpha internal consistency reliability estimate was 0.98 and adjusted (corrected for overlap) item-total correlations ranged from 0.805 to 0.903. For BDI- II, alpha was 0.965 and adjusted item-total correlations ranged from 0.561 to 0.821. For the 36 items, alpha was 0.983 and adjusted item-total correlations ranged from 0.558 to 0.883.

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 Depression	15	0.980	0.805	0.865	0.903
BDI-II	21	0.965	0.561	0.741	0.821
Combined	36	0.983	0.558	0.779	0.883

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 Depression, the fit statistics were as follows: CFI = 0.994, TLI = 0.992, and RMSEA = 0.091. For BDI-II, CFI = 0.978, TLI = 0.976, and RMSEA = 0.075. For the 36 items, CFI = 0.975,

TLI = 0.974, and RMSEA = 0.077. 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 Depression	15	1120	0.994	0.992	0.091
BDI-II	21	1120	0.978	0.976	0.075
Combined	36	1120	0.975	0.974	0.077

5.7.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 36 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 36 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 15 PROMIS Depression items served as anchor items to transform the item parameter estimates for the BDI-II 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 5.7.3 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 Depression items were constrained to their final bank values, while the BDI-II items were calibrated, under the constraints imposed by the anchor items.

Table 5.7.3: IRT Linking Constants

	A	B
Mean/Mean	1.297	0.597
Mean/Sigma	1.349	0.574
Haebara	1.327	0.594
Stocking-Lord	1.338	0.578

The item parameter estimates for the BDI-II items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.7.3. The BDI-II 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 BDI-II 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.

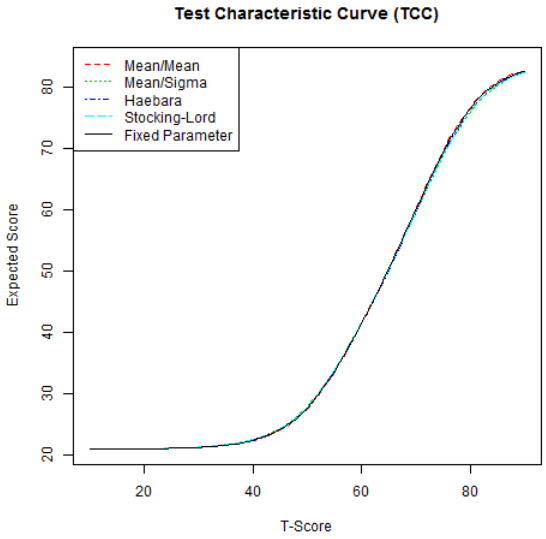


Figure 5.7.5: Test Characteristic Curves (TCC) from Different Linking Methods

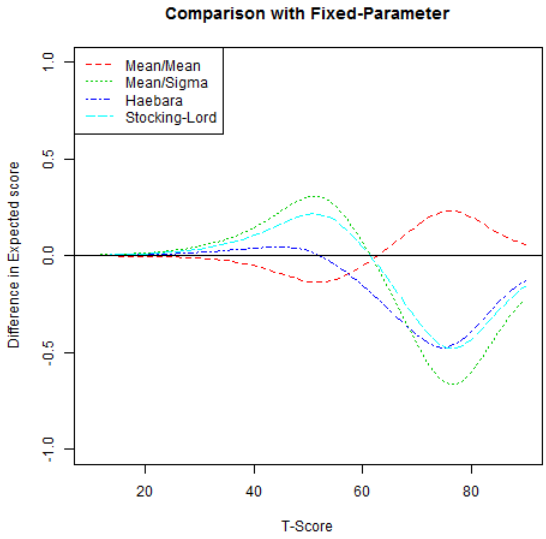


Figure 5.7.6: Difference in Test Characteristic Curves (TCC)

Table 5.7.4 shows the fixed-parameter calibration item parameter estimates for BDI-II. The marginal reliability estimate for BDI-II based on the item parameter estimates was 0.887. The marginal reliability estimates for PROMIS Depression and the combined set were 0.914 and 0.942, respectively. The slope parameter estimates for BDI-II ranged from 1.33 to 3.48 with a mean of 2.27. The slope parameter estimates for PROMIS Depression ranged from 2.38 to 4.45 with a mean of 3.53. We also derived scale information functions based on the fixed- parameter calibration result. Figure 5.7.7 displays the scale information functions for PROMIS Depression, BDI-II, and the combined set of 36. 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 for BDI-II

a	cb1	cb2	cb3	NCAT
2.777	0.639	1.861	2.528	4
2.223	0.212	1.715	2.627	4
2.569	0.443	1.499	2.594	4
2.721	0.197	1.581	2.618	4
2.576	0.552	1.773	2.596	4
2.417	0.857	1.679	2.228	4
2.829	0.575	1.397	2.333	4
2.362	0.428	1.567	2.617	4
2.007	1.269	2.332	3.042	4
2.192	0.823	1.739	2.294	4
2.265	0.633	1.957	2.778	4
2.425	0.470	1.701	2.410	4
2.528	0.651	1.678	2.435	4
3.483	0.708	1.444	2.377	4
1.751	-0.327	1.500	2.887	4
1.330	-0.336	1.664	2.974	4
2.184	0.364	1.648	2.468	4
1.763	0.338	1.911	2.944	4
2.230	0.355	1.486	2.551	4
1.786	-0.040	1.568	2.697	4
1.343	0.259	1.501	2.551	4

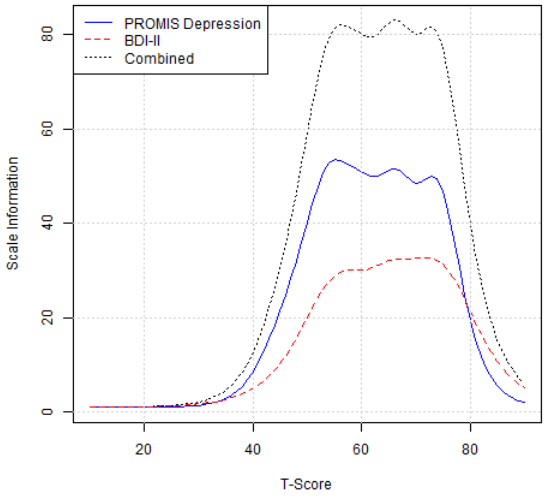


Figure 5.7.7: Comparison of Scale Information Functions

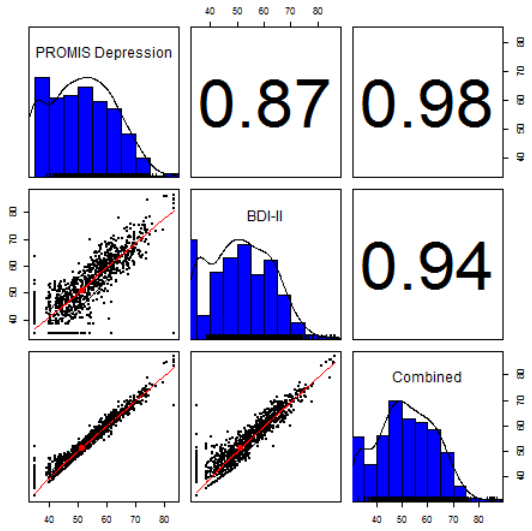


Figure 5.7.8: Comparison of IRT Scaled Scores

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 BDI-II to a scaled score on PROMIS Depression can be useful. Based on the BDI-II 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 BDI-II 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.

5.7.6. Equipercentile Linking

We mapped each raw summed score point on BDI-II to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on BDI-II. 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 BDI-II 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 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).

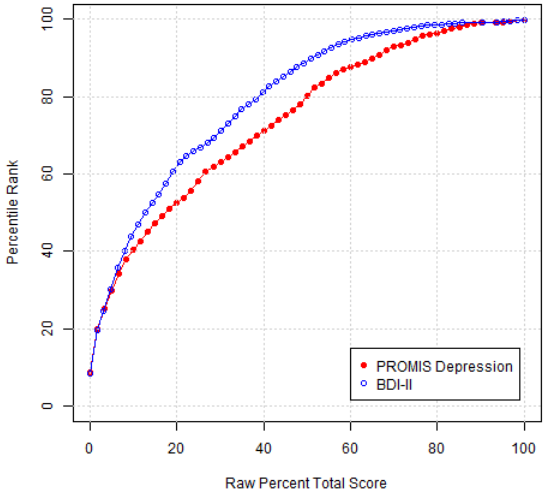


Figure 5.7.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

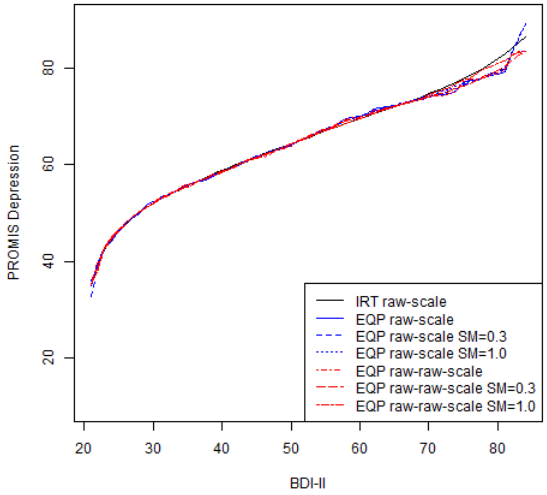


Figure 5.7.10: Equipercentile Linking Functions

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 be- tween the PROMIS Depression

T-scores and BDI-II 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, IRT pattern scoring produced the best result (0.866), followed by EQP raw-scale SM=1.0 (0.859). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.869), followed by EQP raw-scale SM=1.0 (5.932).

Table 5.7.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.866	0.209	5.868	5.869
IRT raw-scale	0.858	0.208	6.011	6.011
EQP raw-scale SM=0.0	0.857	0.154	6.008	6.007
EQP raw-scale SM=0.3	0.851	0.525	6.328	6.347
EQP raw-scale SM=1.0	0.859	0.046	5.934	5.932
EQP raw-raw-scale SM=0.0	0.857	0.137	5.996	5.995
EQP raw-raw-scale SM=0.3	0.857	0.179	6.005	6.005
EQP raw-raw-scale SM=1.0	0.856	0.159	6.018	6.017

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1104) 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 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, 0.647. That is, the difference between the mean PROMIS Depression T-score and the mean equated BDI-II T-score based on a similar sample of 75 cases is expected to be around ±1.29 (i.e., 2 × 0.647).

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	0.221	1.162	0.205	0.812	0.217	0.647
IRT raw-scale	0.205	1.188	0.203	0.840	0.208	0.671
EQP raw-scale SM=0.0	0.139	1.192	0.163	0.835	0.151	0.669
EQP raw-scale SM=0.3	0.528	1.254	0.512	0.868	0.519	0.707
EQP raw-scale SM=1.0	0.053	1.168	0.046	0.832	0.041	0.658
EQP raw-raw-scale SM=0.0	0.136	1.169	0.130	0.829	0.140	0.671
EQP raw-raw-scale SM=0.3	0.158	1.169	0.186	0.839	0.171	0.672
EQP raw-raw-scale SM=1.0	0.165	1.195	0.154	0.839	0.161	0.671

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 Depression and K6

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression item bank (a selection of 20 highly informative items) and K6 (6 items). Both instruments were scaled such that higher scores represent higher levels of Depression. We did not exclude any participants because of missing responses, leaving a final sample of N=748. We created raw summed scores for each of the measures separately and then for the combined. Summing of item scores assumes that all items have positive correlations with the total as examined in the section on Classical Item Analysis. Our sample consisted of 748 participants.

5.8.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 100 for PROMIS Depression and 30 for K6. 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 Depression and K6 was 0.72. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and K6 was 0.76. The correlations between the combined score and the measures were 0.99 and 0.82 for PROMIS Depression and K6, respectively.

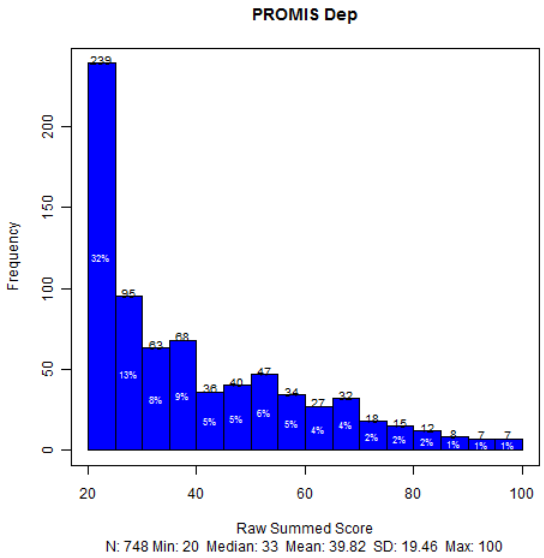


Figure 5.8.1: Raw Summed Score Distribution - PROMIS Depression

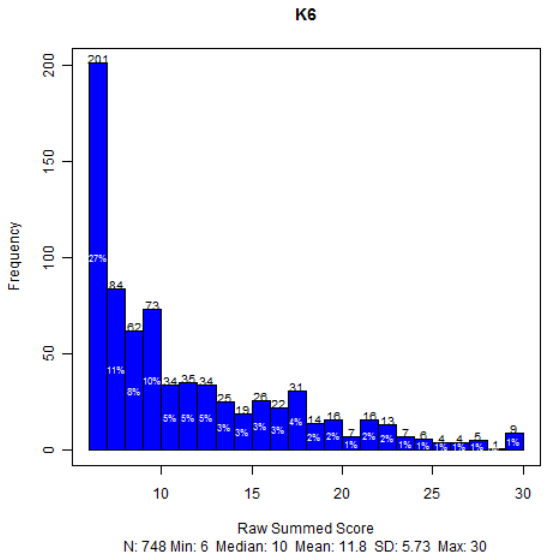


Figure 5.8.2: Raw Summed Score Distribution – K6

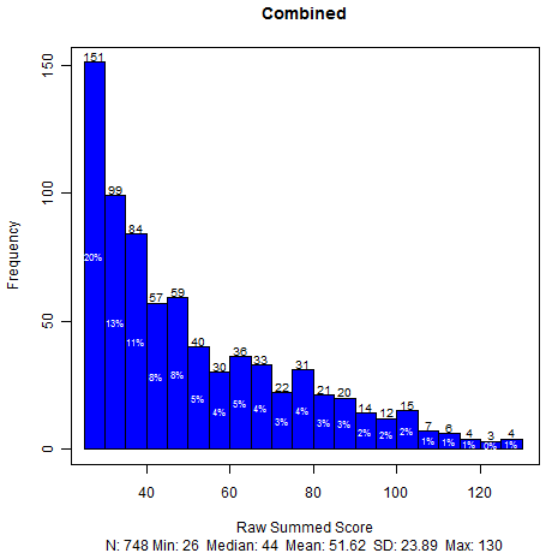


Figure 5.8.3: Raw Summed Score Distribution – Combined

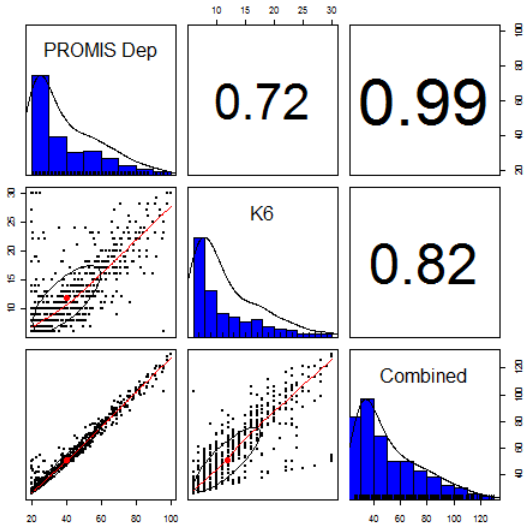


Figure 5.8.4: Scatter Plot Matrix of Raw Summed Scores

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 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 K6, alpha was 0.897 and adjusted item-total correlations ranged from 0.629 to 0.812. For the 26 items, alpha was 0.977 and adjusted item-total correlations ranged from 0.474 to 0.88.

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 Depression	20	0.979	0.741	0.826	0.880
K6	6	0.897	0.629	0.723	0.812
Combined	26	0.977	0.474	0.777	0.880

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 Depression, the fit statistics were as follows: CFI = 0.988, TLI= 0.986, and RMSEA = 0.089. For K6, CFI = 0.979, TLI = 0.965, and RMSEA = 0.166. For the 26 items, CFI = 0.96, TLI = 0.956, and RMSEA = 0.125.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 Dep	20	748	0.988	0.986	0.089
K6	6	748	0.979	0.965	0.166
Combined	26	748	0.960	0.956	0.125

5.8.4. 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 Depression items served as anchor items to transform the item parameter estimates for the K6 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 5.8.3 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 Depression items were constrained to their final bank values, while the K6 items were calibrated, under the constraints imposed by the anchor items.

Table 5.8.3: IRT Linking Constants

	A	B
Mean/Mean	1.165	0.381
Mean/Sigma	1.229	0.337
Haebara	1.231	0.364
Stocking-Lord	1.217	0.349

The item parameter estimates for the K6 items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.8.3. The K6 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 K6 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.

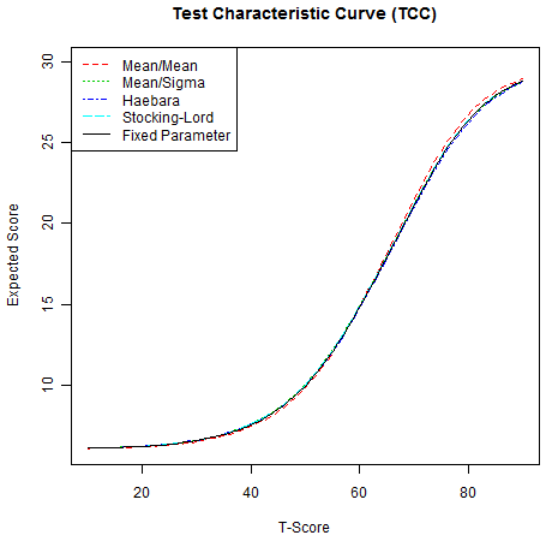


Figure 5.8.5: Test Characteristic Curves (TCC) from Different Linking Methods

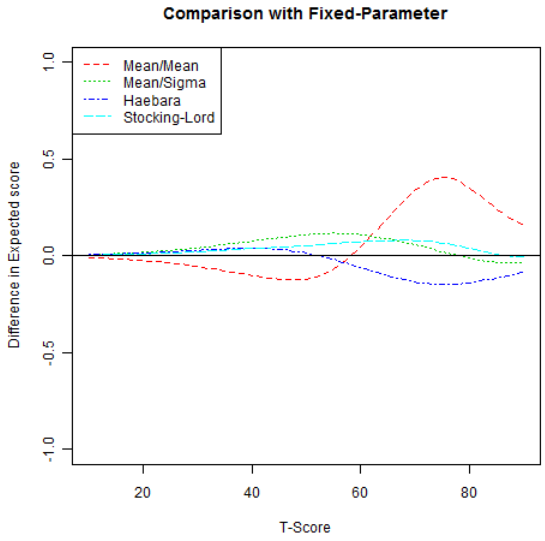


Figure 5.8.6: Difference in Test Characteristic Curves (TCC)

Table 5.8.4 shows the fixed-parameter calibration item parameter estimates for K6. The marginal reliability estimate for K6 based on the item parameter estimates was 0.747. The marginal reliability estimates for PROMIS Depression and the combined set were 0.929 and 0.941, respectively. The slope parameter estimates for K6 ranged from 1.03 to 2.71 with a mean of 1.73. 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 5.8.7 displays the scale information functions for PROMIS Depression, 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 5.8.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.8.4: Fixed-Parameter Calibration Item Parameter Estimates for K6

a	cb1	cb2	cb3	cb4	NCAT
1.029	-0.741	0.936	2.145	2.884	5
1.777	0.291	1.182	1.960	2.463	5
1.259	-0.293	1.029	2.162	3.157	5
2.086	0.611	1.375	2.004	2.592	5
1.553	-0.348	0.879	1.895	2.858	5
2.706	0.562	1.228	1.868	2.530	5

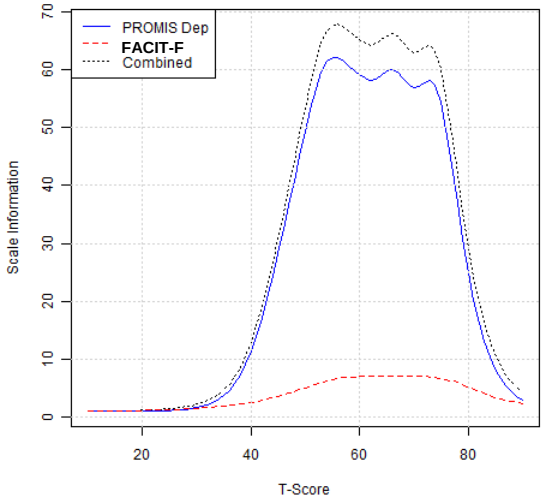


Figure 5.8.7: Comparison of Scale Information Functions

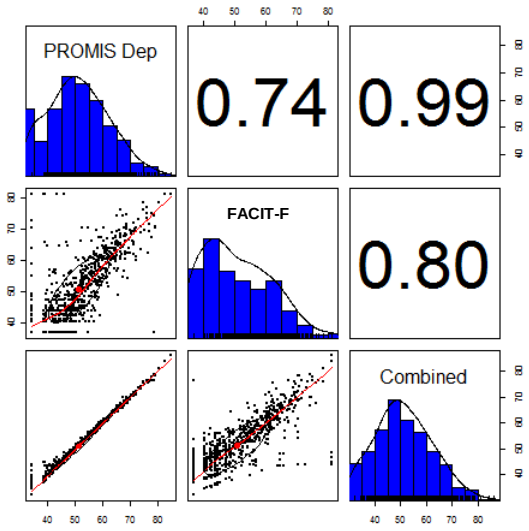


Figure 5.8.8: Comparison of IRT Scaled Scores

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 K6 to a scaled score on PROMIS Depression 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 22 can be used to map simple raw summed scores from K6 to T-score values linked to the PROMIS Depression metric. Each raw summed score point and corresponding PROMIS Depression 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 K6 to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression 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 5.8.9 displays the cumulative distribution functions of the measures. Figure 5.8.10 shows the equipercentile linking functions based on raw summed scores, from K6 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 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).

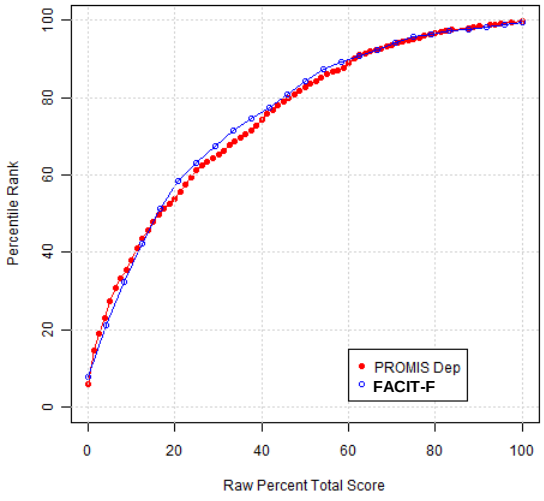


Figure 5.8.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

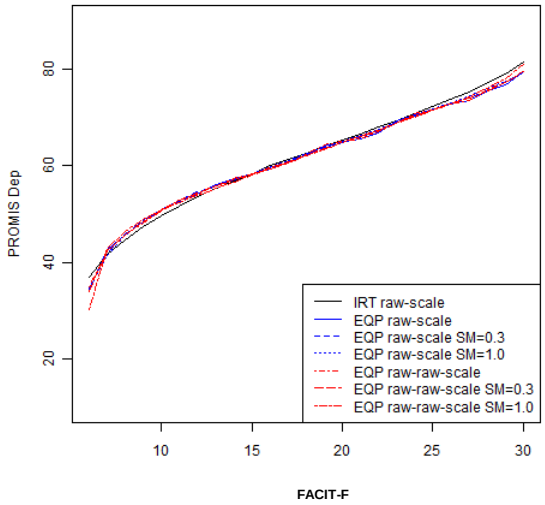


Figure 5.8.10: Equipercentile Linking Functions

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 Depression 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 5.8.10), 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 (0.741), followed by EQP raw-scale SM=1.0 (0.69). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (7.852), followed by EQP raw-raw-scale SM=0.0 (8.565)

Table 5.8.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.741	0.126	7.856	7.852
IRT raw-scale	0.686	0.154	8.595	8.591
EQP raw-scale SM=0.0	0.688	0.157	8.626	8.622
EQP raw-scale SM=0.3	0.689	0.128	8.613	8.608
EQP raw-scale SM=1.0	0.690	0.140	8.599	8.594
EQP raw-raw-scale SM=0.0	0.688	0.145	8.570	8.565
EQP raw-raw-scale SM=0.3	0.689	0.271	8.645	8.643
EQP raw-raw-scale SM=1.0	0.682	0.695	9.096	9.117

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

Table 5.8.6 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 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, 0.869. That is, the difference between the mean PROMIS Depression T-score and the mean equated K6 T-score based on a similar sample of 75 cases is expected to be around ±1.74 (i.e., 2 × 0.869).

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	0.117	1.530	0.141	1.080	0.126	0.869
IRT raw-scale	0.149	1.681	0.175	1.195	0.171	0.927
EQP raw-scale SM=0.0	0.140	1.713	0.166	1.178	0.160	0.950
EQP raw-scale SM=0.3	0.119	1.701	0.110	1.187	0.114	0.945
EQP raw-scale SM=1.0	0.162	1.693	0.133	1.164	0.145	0.949
EQP raw-raw-scale SM=0.0	0.120	1.678	0.128	1.174	0.151	0.939
EQP raw-raw-scale SM=0.3	0.278	1.727	0.283	1.160	0.272	0.948
EQP raw-raw-scale SM=1.0	0.717	1.807	0.704	1.255	0.691	0.993

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 Depression and PANAS Negative Affect

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression item bank (a selection of 15 highly informative items) and PANAS Negative Affect (10 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. Our sample consisted of 1,120 participants (N = 1,105 for participants with complete responses).

5.9.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 75 for PROMIS Depression and 50 for PANAS NA. 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 Depression and PANAS NA was 0.85. The disattenuated (corrected for unreliabilities) correlation between PROMIS Depression and PANAS NA was 0.88. The correlations between the combined score and the measures were 0.98 and 0.94 for PROMIS Depression and PANAS NA, respectively.

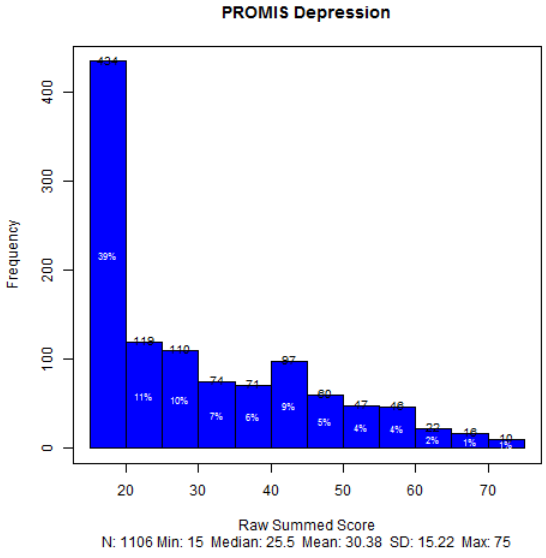


Figure 5.9.1: Raw Summed Score Distribution - PROMIS Depression

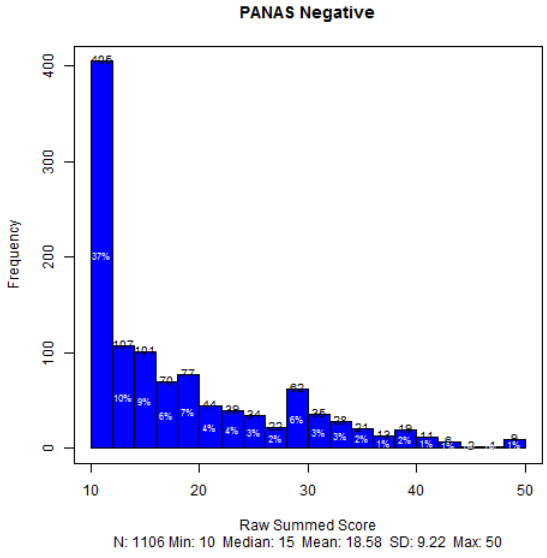


Figure 5.9.2: Raw Summed Score Distribution – PANAS NA

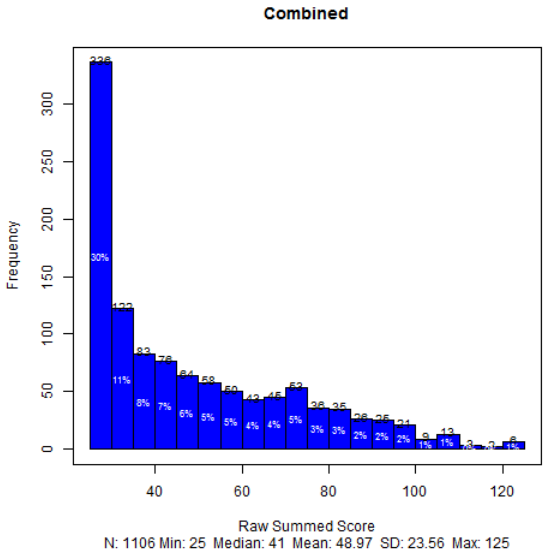


Figure 5.9.3: Raw Summed Score Distribution – Combined

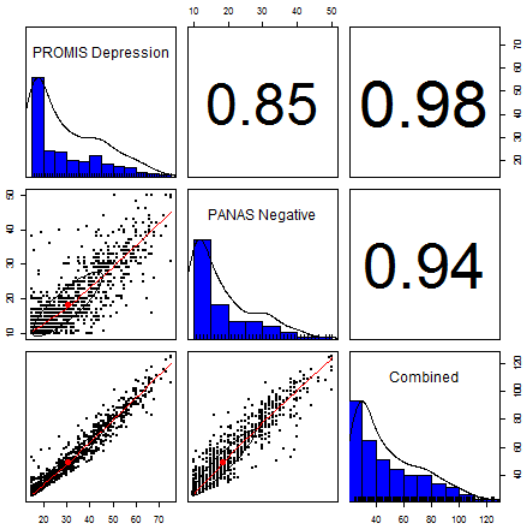


Figure 5.9.4: Scatter Plot Matrix of Raw Summed Scores

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 Depression, Cronbach's alpha internal consistency reliability estimate was 0.98 and adjusted (corrected for overlap) item-total correlations ranged from 0.805 to 0.903. For PANAS NA, alpha was 0.954 and adjusted item-total correlations ranged from 0.74 to 0.844. For the 25 items, alpha was 0.982 and adjusted item-total correlations ranged from 0.686 to 0.888.

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 Depression	15	0.980	0.805	0.865	0.903
PANAS NA	10	0.954	0.740	0.801	0.844
Combined	25	0.982	0.686	0.819	0.888

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 Depression, the fit statistics were as follows: CFI = 0.994, TLI = 0.992, and RMSEA = 0.091. For PANAS NA, CFI = 0.984, TLI = 0.98, and RMSEA= 0.125. For the 25 items, CFI = 0.978, TLI = 0.976, and RMSEA = 0.102. 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 Depression	15	1120	0.994	0.992	0.091
PANAS NA	10	1120	0.984	0.980	0.125
Combined	25	1120	0.978	0.976	0.102

5.9.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 25 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 25 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 15 PROMIS Depression items served as anchor items to transform the item parameter estimates for the PANAS NA 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 5.9.3 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 Depression items were constrained to their final bank values, while the PANAS NA items were calibrated, under the constraints imposed by the anchor items.

Table 5.9.3: IRT Linking Constants

	A	B
Mean/Mean	1.276	0.442
Mean/Sigma	1.281	0.439
Haebara	1.265	0.458
Stocking-Lord	1.276	0.443

The item parameter estimates for the PANAS NA items were linked to the PROMIS Depression metric using the transformation constants shown in Table 5.9.3. The PANAS NA 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 PANAS NA 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.

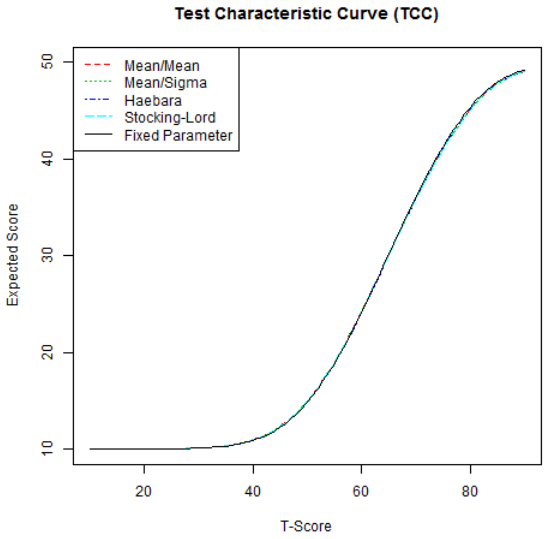


Figure 5.9.5: Test Characteristic Curves (TCC) from Different Linking Methods

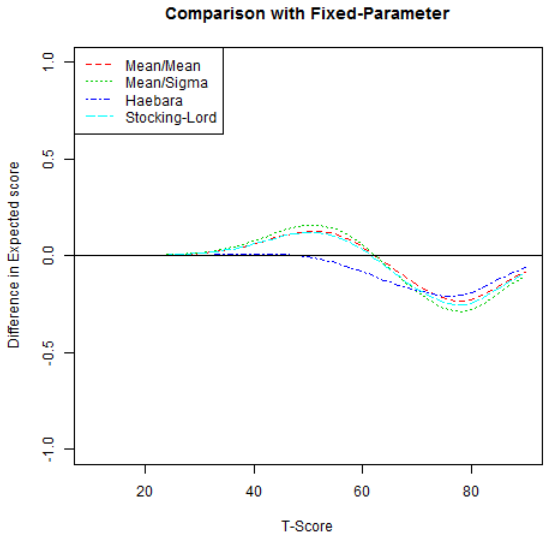


Figure 5.9.6: Difference in Test Characteristic Curves (TCC)

Table 5.9.4 shows the fixed-parameter calibration item parameter estimates for PANAS NA. The marginal reliability estimate for PANAS NA based on the item parameter estimates was 0.852. The marginal reliability estimates for PROMIS Depression and the combined set were 0.914 and 0.935, respectively. The slope parameter estimates for PANAS NA ranged from 1.75 to 2.55 with a mean of 2.28. The slope parameter estimates for PROMIS Depression ranged from 2.38 to 4.45 with a mean of 3.53. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.9.7 displays the scale information functions for PROMIS Depression, PANAS NA, and the combined set of 25. 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 for PANAS NA

a	cb1	cb2	cb3	cb4	NCAT
2.443	-0.072	0.912	1.742	2.792	5
2.248	-0.262	0.867	1.714	2.898	5
2.331	0.445	1.158	1.941	2.844	5
2.530	0.516	1.258	1.923	2.733	5
1.754	0.524	1.485	2.303	3.413	5
2.008	-0.417	0.877	1.699	2.662	5
2.441	0.619	1.263	1.862	2.587	5
2.294	0.026	0.944	1.608	2.473	5
2.227	0.330	1.171	1.853	2.692	5
2.548	0.544	1.233	1.797	2.610	5

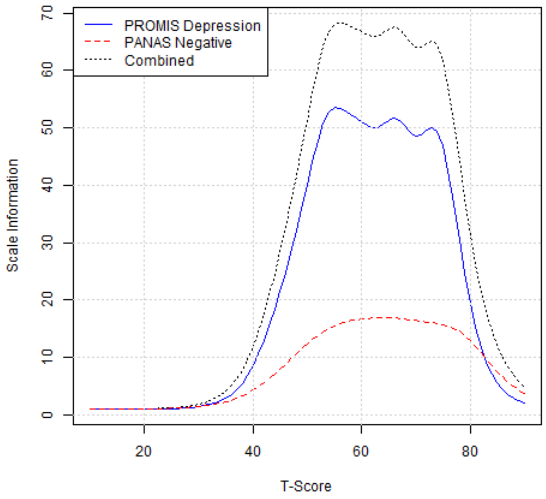


Figure 5.9.7: Comparison of Scale Information Functions

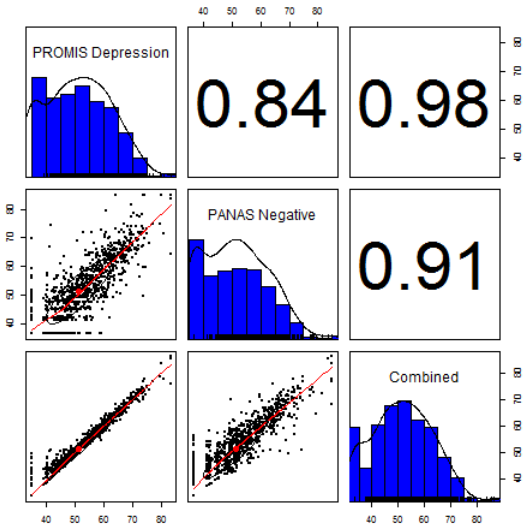


Figure 5.9.8: Comparison of IRT Scaled Scores

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 PANAS NA to a scaled score on PROMIS Depression can be useful. Based on the PANAS NA 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 PANAS NA 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.

5.9.6. Equipercentile Linking

We mapped each raw summed score point on PANAS NA to a correspond- ing scaled score on PROMIS Depression by identifying scores on PROMIS De- pression that have the same percentile ranks as scores on PANAS NA. 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 PANAS NA 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 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).

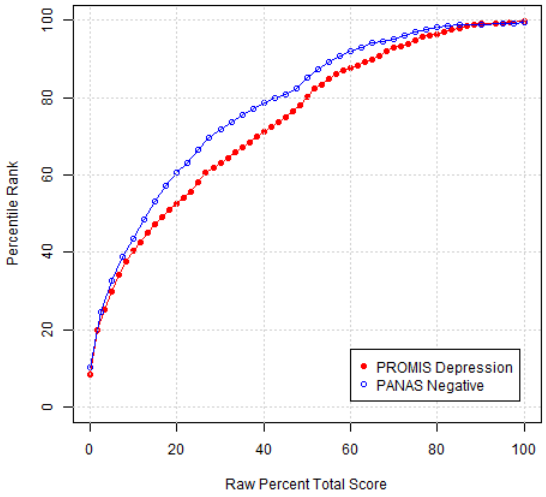


Figure 5.9.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

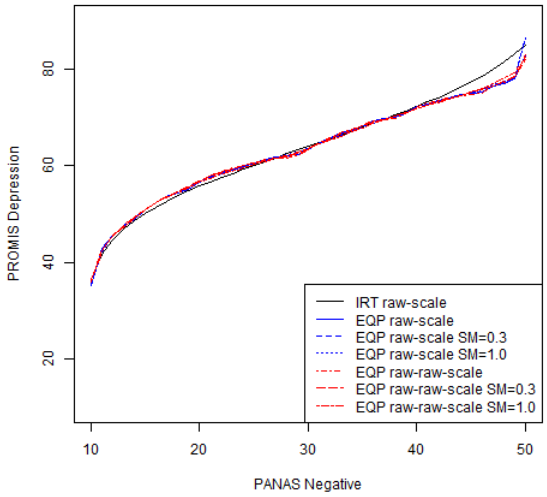


Figure 5.9.10: Equipercentile Linking Functions

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 Depression T-

scores and PANAS NA 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 raw-scale produced the best result (0.841), followed by IRT pattern scoring (0.841). 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 (6.267), followed by IRT raw- scale (6.31).

Table 5.9.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.841	0.193	6.319	6.320
IRT raw-scale	0.841	0.378	6.301	6.310
EQP raw-scale SM=0.0	0.839	0.227	6.404	6.405
EQP raw-scale SM=0.3	0.839	0.205	6.406	6.406
EQP raw-scale SM=1.0	0.840	0.188	6.381	6.381
EQP raw-raw-scale SM=0.0	0.840	0.156	6.336	6.335
EQP raw-raw-scale SM=0.3	0.840	0.039	6.270	6.267
EQP raw-raw-scale SM=1.0	0.840	0.163	6.314	6.314

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1106) 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 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, EQP raw-raw-scale SM=0.3 produced the smallest standard error, 0.695. That is, the difference between the mean PROMIS Depression T-score and the mean equated PANAS NA T-score based on a similar sample of 75 cases is expected to be around ±1.39 (i.e., 2 × 0.695).

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	0.187	1.258	0.190	0.868	0.183	0.705
IRT raw-scale	0.351	1.238	0.388	0.869	0.386	0.714
EQP raw-scale SM=0.0	0.233	1.267	0.212	0.877	0.238	0.714
EQP raw-scale SM=0.3	0.208	1.266	0.202	0.899	0.205	0.705
EQP raw-scale SM=1.0	0.164	1.255	0.190	0.877	0.190	0.714
EQP raw-raw-scale SM=0.0	0.147	1.254	0.160	0.874	0.154	0.708
EQP raw-raw-scale SM=0.3	0.031	1.243	0.036	0.871	0.038	0.695
EQP raw-raw-scale SM=1.0	0.157	1.248	0.167	0.876	0.167	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 25) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.10. PROMIS Depression and PHQ-2

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

5.10.1. Raw Summed Score Distribution

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

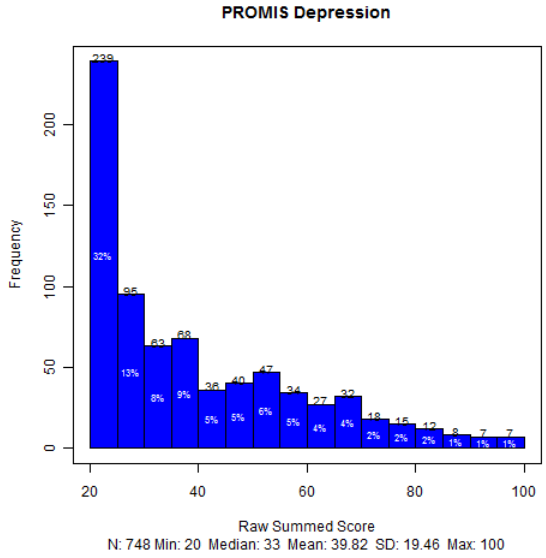


Figure 5.10.1: Raw Summed Score Distribution - PROMIS Depression

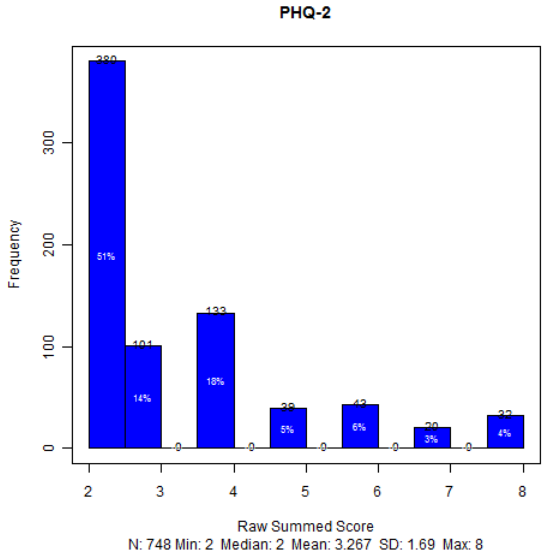


Figure 5.10.2: Raw Summed Score Distribution – PHQ-2

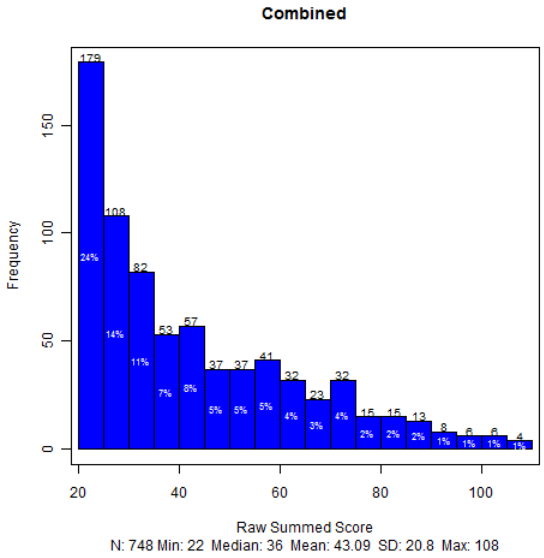


Figure 5.10.3: Raw Summed Score Distribution – Combined

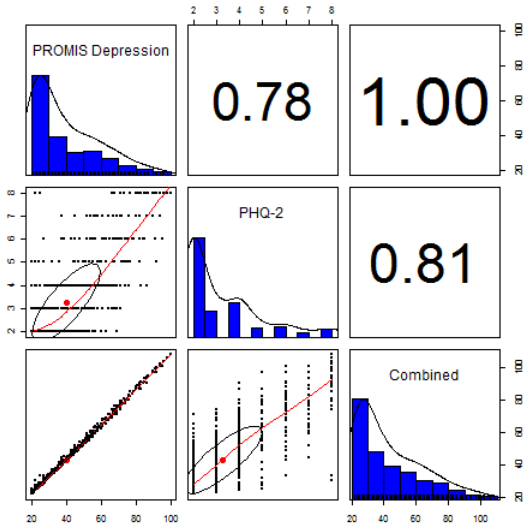


Figure 5.10.4: Scatter Plot Matrix of Raw Summed Scores

5.10.2. Classical Item Analysis

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

Table 5.10.1: Classical Item Analysis

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

5.10.3. Confirmatory Factor Analysis (CFA)

To assess the dimensionality of the measures, a categorical confirmatory factor analysis (CFA) was carried out using the WLSMV estimator of Mplus on a subset of cases without missing responses. A single factor model (based on polychoric correlations) was run on PROMIS Depression and on the combined item set. Table 5.10.2 summarizes the model fit statistics.

Table 5.10.2: CFA Fit Statistics

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

5.10.4. Item Response Theory (IRT) Linking

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

Table 5.10.3: IRT Linking Constants

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

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

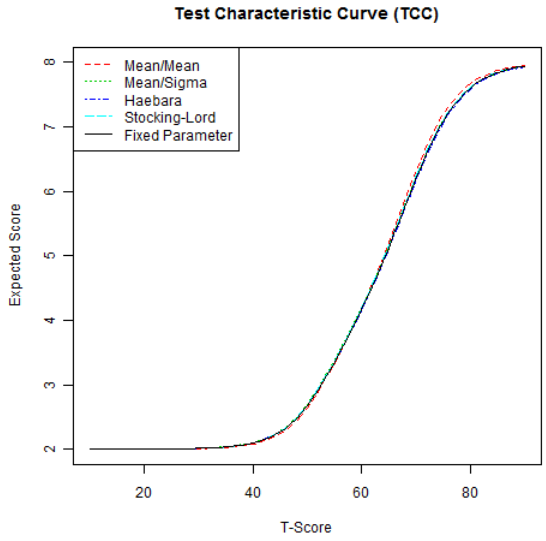


Figure 5.10.5: Test Characteristic Curves (TCC) from Different Linking Methods

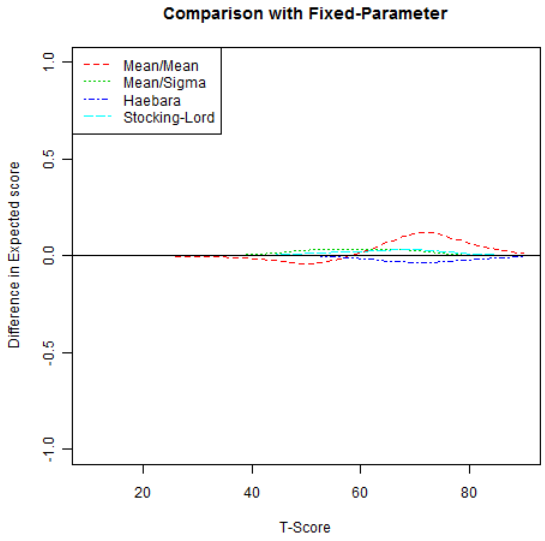


Figure 5.10.6: Difference in Test Characteristic Curves (TCC)

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

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

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

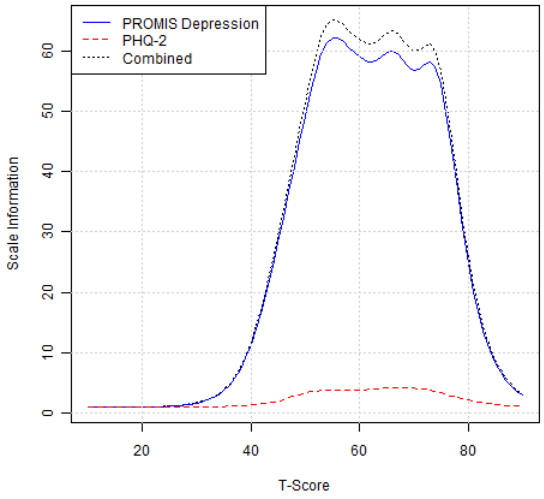


Figure 5.10.7: Comparison of Scale Information Functions

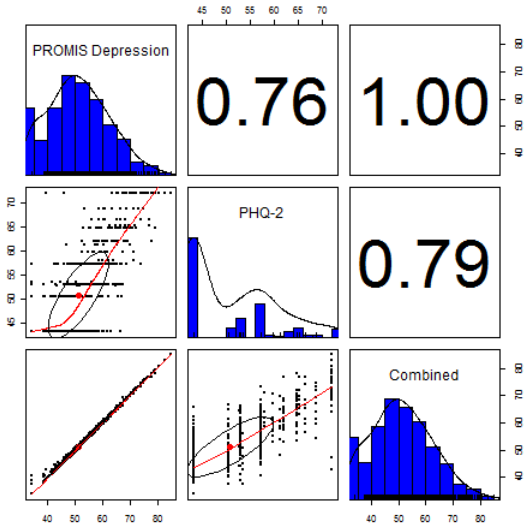


Figure 5.10.8: Comparison of IRT Scaled Scores

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

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

5.10.6. Equipercentile Linking

We mapped each raw summed score point on PHQ-2 to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on PHQ-2. Theoretically, the equipercentile linking function is symmetrical for continuous random variables (X and Y). Therefore, the linking function for the values in X to those in Y is the same as that for the values in Y to those in X. However, for discrete variables like raw summed scores the equipercentile linking functions can be slightly different (due to rounding errors and differences in score ranges) and hence may need to be obtained separately. Figure 5.10.9 displays the cumulative distribution functions of the measures. Figure 5.10.10 shows the equipercentile linking functions based on raw summed scores, from PHQ-2 to PROMIS Depression. When the number of raw summed score points differs substantially, the equipercentile linking functions could deviate from each other

noticeably. The problem can be exacerbated when the sample size is small. Appendix Table 29 and Appendix Table 30 show the equipercentile crosswalk tables. The result shown in Appendix Table 29 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 30 shows the result based on the indirect (raw summed score to raw summed score equivalent to scaled score equivalent) approach (Refer to Section 4.2 for details). Three separate equipercentile equivalents are presented: one is equipercentile without post smoothing ("Equipercentile Scale Score Equivalents") and two with different levels of postsmoothing, i.e., "Equipercentile Equivalents with Postsmoothing (Less Smoothing)" and "Equipercentile Equivalents with Postsmoothing (More Smoothing)." Postsmoothing values of 0.3 and 1.0 were used for "Less" and "More," respectively (Refer to Brennan, 2004 for details).

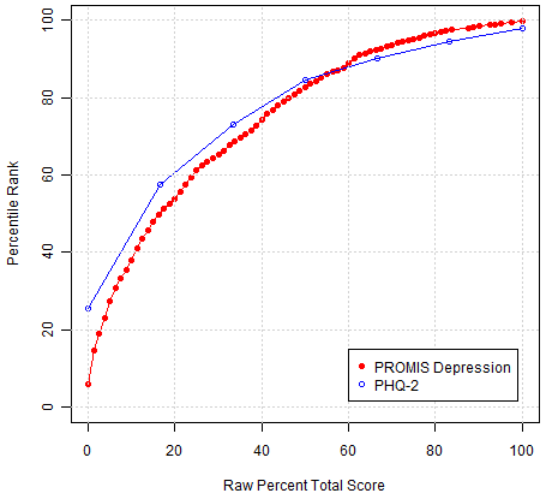


Figure 5.10.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

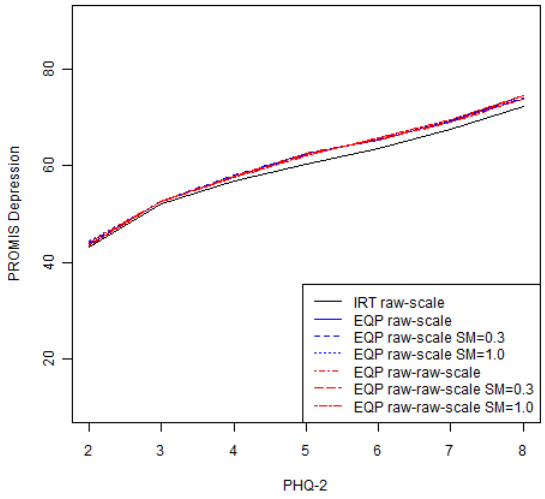


Figure 5.10.10: Equipercentile Linking Functions

5.10.7. Summary and Discussion

The purpose of linking is to establish the relationship between scores on two measures of closely related traits. The relationship can vary across linking methods and samples employed. In equipercentile linking, the relationship is determined based on the distributions of scores in a given sample. Although IRT-based linking can potentially offer sample-invariant results, they are based on estimates of item parameters, and hence subject to sampling errors. A potential issue with IRT-based linking methods is, however, the violation of model assumptions as a result of combining items from two measures (e.g., unidimensionality and local independence). As displayed in Figure 5.10.10, the relationships derived from various linking methods are consistent, which suggests that a robust linking relationship can be determined based on the given sample.

To further facilitate the comparison of the linking methods, Table 5.10.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Depression T-

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

Table 5.10.5: Observed vs. Linked T-scores

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

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

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

Table 5.10.6: Comparison of Resampling Results

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

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

5.11. PROMIS Fatigue and Neuro-QoL Fatigue

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 (a selection of 14 items) and Neuro-QoL Fatigue (19 items). PROMIS Fatigue was scaled such that higher scores represent higher 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. Our sample consisted of 1,120 participants (N = 1,114 for participants with complete responses).

5.11.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 70 for PROMIS Fatigue and 95 for Neuro-QoL Fatigue. 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 Fatigue and Neuro-QoL Fatigue was 0.95. The disattenuated (corrected for unreliabilities) correlation between PROMIS Fatigue and Neuro-QoL Fatigue was 0.97. The correlations between the combined score and the measures were 0.98 and 0.99 for PROMIS Fatigue and Neuro-QoL Fatigue, respectively.

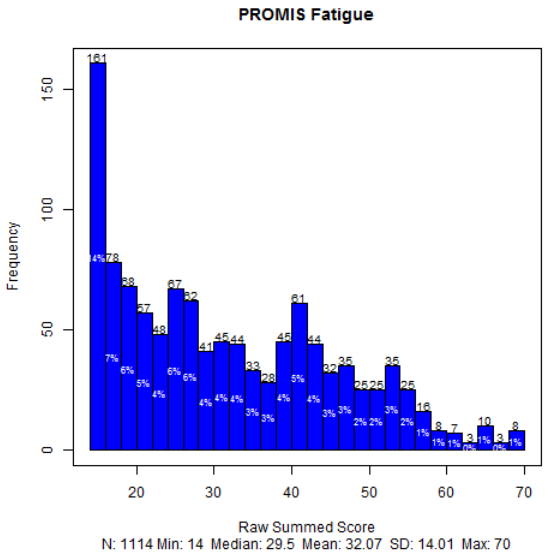


Figure 5.11.1: Raw Summed Score Distribution - PROMIS Fatigue

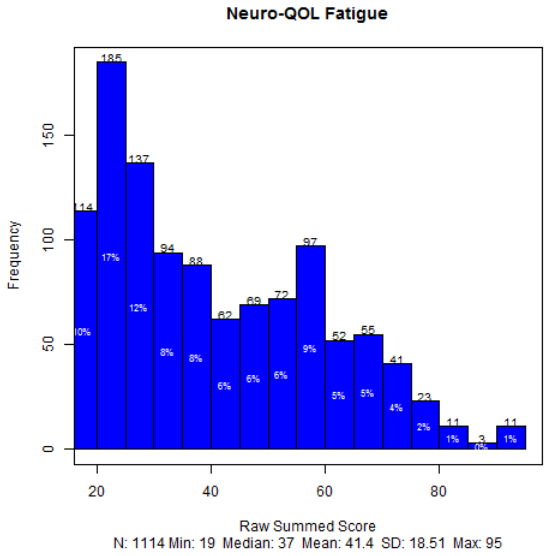


Figure 5.11.2: Raw Summed Score Distribution – Neuro-QOL Fatigue

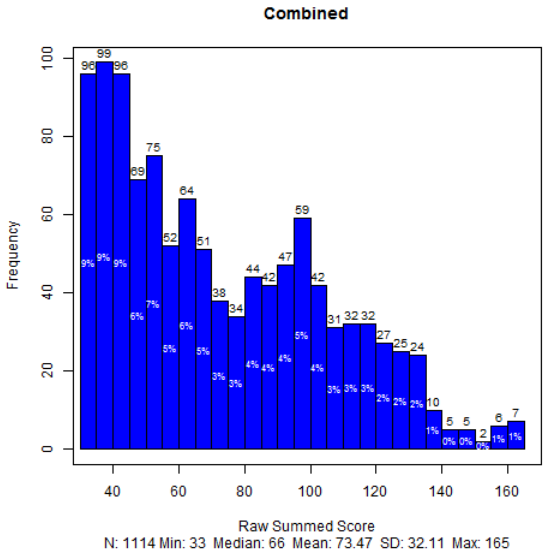


Figure 5.11.3: Raw Summed Score Distribution – Combined

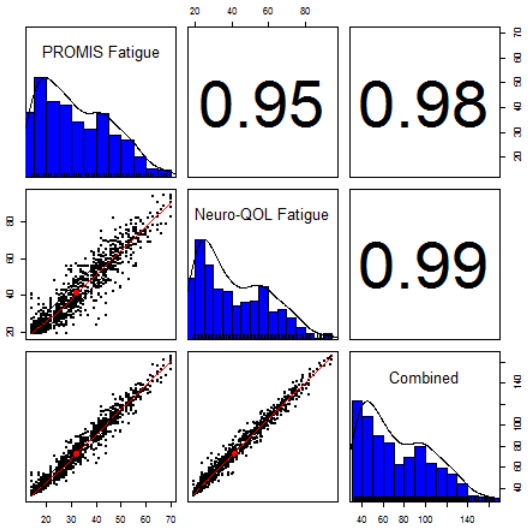


Figure 5.11.4: Scatter Plot Matrix of Raw Summed Scores

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 Fatigue, Cronbach's alpha internal consistency reliability estimate was 0.974 and adjusted (corrected for overlap) item-total correlations ranged from 0.735 to 0.882. For Neuro-QoL Fatigue, alpha was 0.978 and adjusted item-total correlations ranged from 0.747 to 0.877. For the 33 items, alpha was 0.987 and adjusted item-total correlations ranged from 0.742 to 0.888.

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 Fatigue	14	0.974	0.735	0.843	0.882
Neuro-QoL Fatigue	19	0.978	0.747	0.830	0.877
Combined	33	0.987	0.742	0.834	0.888

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 Fatigue, the fit statistics were as follows: CFI = 0.989, TLI = 0.987, and RMSEA = 0.115. For Neuro-QoL Fatigue, CFI = 0.974, TLI = 0.971, and RMSEA = 0.14. For the 33 items, CFI = 0.973, TLI = 0.971, and RMSEA = 0.105. 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 Fatigue	14	1120	0.989	0.987	0.115
Neuro-QoL Fatigue	19	1120	0.974	0.971	0.140
Combined	33	1120	0.973	0.971	0.105

5.11.4. 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 14 PROMIS Fatigue items served as anchor items to transform the item parameter estimates for the Neuro-QoL Fatigue 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 5.11.3 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 Fatigue items were constrained to their final bank values, while the Neuro-QoL Fatigue items were calibrated, under the constraints imposed by the anchor items.

Table 5.11.3: IRT Linking Constants

	A	B
Mean/Mean	1.092	0.449
Mean/Sigma	1.229	0.396
Haebara	1.192	0.390
Stocking-Lord	1.193	0.395

The item parameter estimates for the Neuro-QoL Fatigue items were linked to the PROMIS Fatigue metric using the transformation constants shown in Table 5.11.3. The Neuro-QoL Fatigue 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 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.

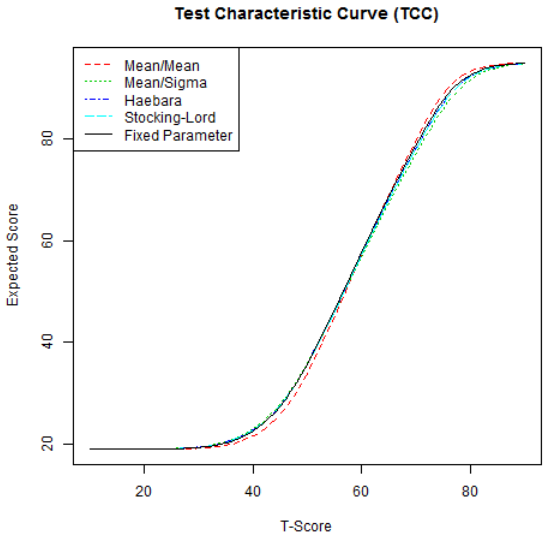


Figure 5.11.5: Test Characteristic Curves (TCC) from Different Linking Methods

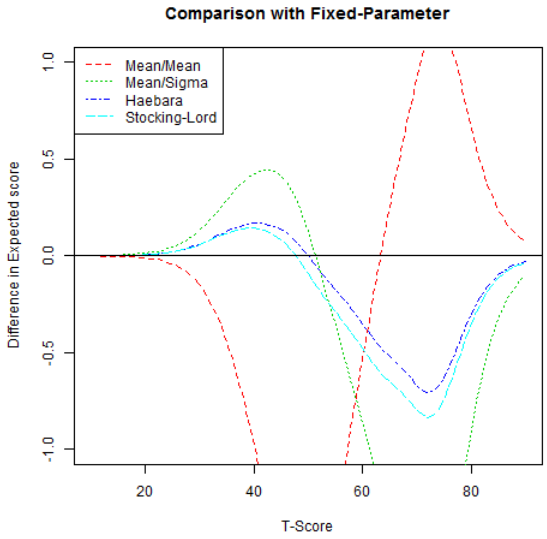


Figure 5.11.6: Difference in Test Characteristic Curves (TCC)

Table 5.11.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QoL Fatigue. The marginal reliability estimate for Neuro-QoL Fatigue based on the item parameter estimates was 0.954. The marginal reliability estimates for PROMIS Fatigue and the combined set were 0.957 and 0.975, respectively. The slope parameter estimates for Neuro-QoL Fatigue ranged from 2.04 to 3.84 with a mean of 3.12. The slope parameter estimates for PROMIS Fatigue ranged from 2.11 to 4.77 with a mean of 3.58. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.11.7 displays the scale information functions for PROMIS Fatigue, Neuro-QoL Fatigue, and the combined set of 33. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.11.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.11.4: Fixed-Parameter Calibration Item Parameter Estimates for Neuro-QoL Fatigue

a	cb1	cb2	cb3	cb4	NCAT
2.992	-0.884	0.097	1.109	2.157	5
3.389	-0.750	0.181	1.116	2.130	5
3.255	-1.068	0.001	1.029	2.208	5
3.275	-0.477	0.328	1.246	2.298	5
3.124	-0.015	0.678	1.508	2.470	5
3.838	-0.229	0.398	1.134	1.947	5
2.877	-1.364	-0.304	0.927	2.042	5
3.605	-0.245	0.530	1.344	2.195	5
3.140	0.136	0.774	1.580	2.434	5
2.038	-0.389	0.454	1.449	2.407	5
3.008	-0.321	0.525	1.442	2.504	5
3.301	-0.546	0.348	1.252	2.158	5
2.995	-0.040	0.550	1.314	2.232	5
2.376	0.333	1.089	1.960	2.781	5
2.767	-0.379	0.418	1.375	2.313	5
3.487	0.035	0.686	1.473	2.290	5
3.090	0.349	0.914	1.655	2.474	5
3.410	0.306	0.864	1.566	2.370	5
3.311	0.266	0.813	1.494	2.266	5

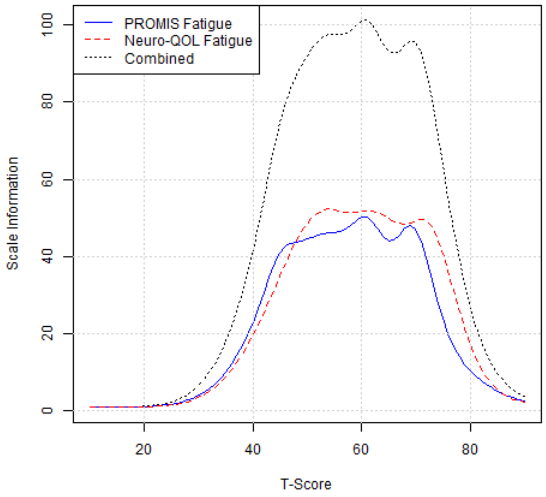


Figure 5.11.7: Comparison of Scale Information Functions

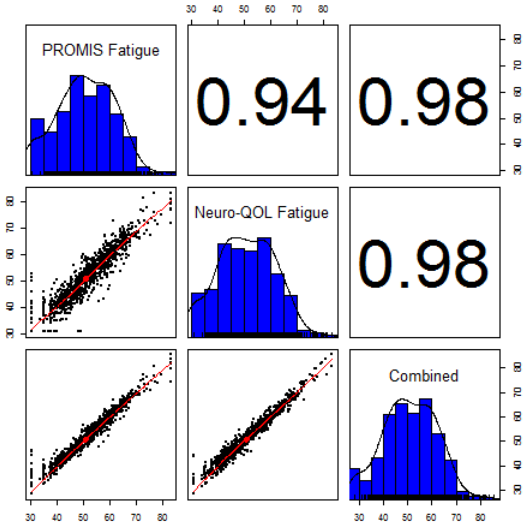


Figure 5.11.8: Comparison of IRT Scaled Scores

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 Neuro-QoL Fatigue to a scaled score on PROMIS Fatigue can be useful. Based on the Neuro-QoL Fatigue 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 Neuro-QOL Fatigue 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.

5.11.6. Equipercentile Linking

We mapped each raw summed score point on Neuro-QoL Fatigue to a corresponding scaled score on PROMIS Fatigue by identifying scores on PROMIS Fatigue that have the same percentile ranks as scores on Neuro-QoL Fatigue. 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 Neuro-QoL Fatigue 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 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).

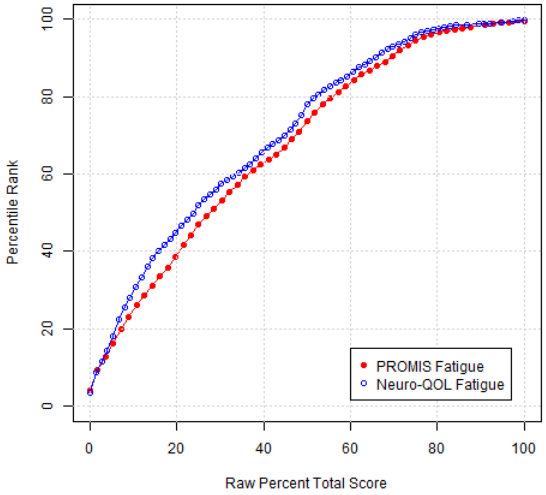


Figure 5.11.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

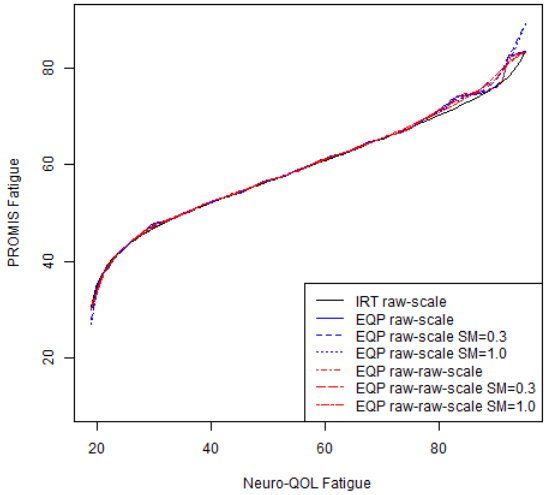


Figure 5.11.10: Equipercentile Linking Functions

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 Fatigue T-

scores and Neuro-QoL Fatigue 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 pattern scoring produced the best result (0.937), followed by EQP raw-raw-scale SM=0.0 (0.937). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (3.883), followed by IRT raw-scale (3.889).

Table 5.11.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.937	-0.026	3.885	3.883
IRT raw-scale	0.937	0.018	3.890	3.889
EQP raw-scale SM=0.0	0.937	-0.016	3.942	3.941
EQP raw-scale SM=0.3	0.934	0.110	4.100	4.099
EQP raw-scale SM=1.0	0.933	0.178	4.166	4.167
EQP raw-raw-scale SM=0.0	0.937	-0.040	3.924	3.923
EQP raw-raw-scale SM=0.3	0.937	-0.031	3.939	3.937
EQP raw-raw-scale SM=1.0	0.937	-0.038	3.939	3.937

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1114) 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 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, 0.431. That is, the difference between the mean PROMIS Fatigue T-score and the mean equated Neuro-QoL Fatigue T-score based on a similar sample of 75 cases is expected to be around ±0.86 (i.e., 2 × 0.431).

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	-0.015	0.772	-0.022	0.534	-0.022	0.431
IRT raw-scale	0.013	0.766	0.018	0.541	0.013	0.439
EQP raw-scale SM=0.0	-0.019	0.777	-0.016	0.546	-0.018	0.436
EQP raw-scale SM=0.3	0.115	0.809	0.118	0.573	0.105	0.455
EQP raw-scale SM=1.0	0.167	0.822	0.178	0.579	0.175	0.474
EQP raw-raw-scale SM=0.0	-0.056	0.775	-0.047	0.539	-0.039	0.437
EQP raw-raw-scale SM=0.3	-0.046	0.785	-0.030	0.542	-0.031	0.439
EQP raw-raw-scale SM=1.0	-0.040	0.778	-0.041	0.540	-0.041	0.436

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 Global Health - Mental and VR-12 - Mental

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Global Health-Mental component, namely the PROMIS Global Mental Health (4 items) and VR12 Mental (6 items). Both instruments were scaled such that higher scores represent higher levels of Global Health-Mental. We excluded 1 participant because of missing responses, leaving a final sample of N=2017. 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 20 for PROMIS Global Mental Health and 33 for VR12 Mental. 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 Global Mental Health and VR12 Mental was 0.69. The disattenuated (corrected for unreliabilities) correlation between PROMIS Global Mental Health and VR12 Mental was 0.85. The correlations between the combined score and the measures were 0.88 and 0.95 for PROMIS Global Mental Health and VR12 Mental, respectively.

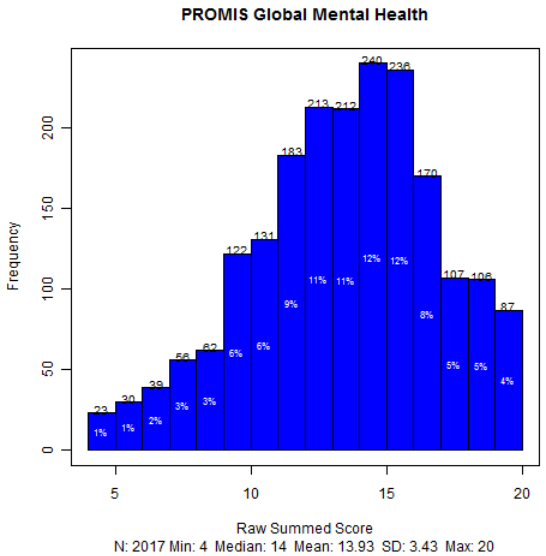


Figure 5.12.1: Raw Summed Score Distribution - PROMIS Global Health – Mental component

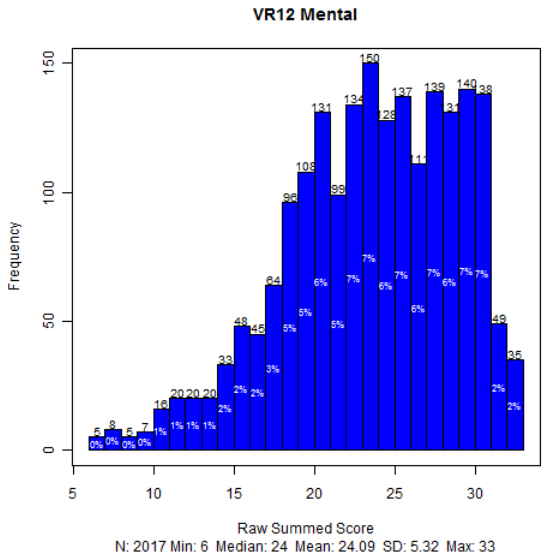


Figure 5.12.2: Raw Summed Score Distribution – VR-12 – Mental component

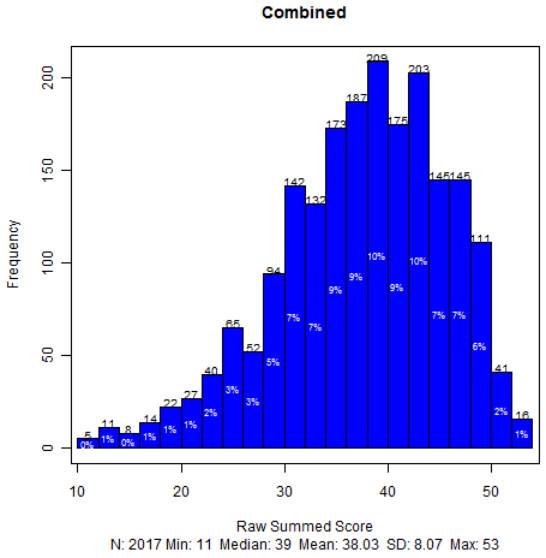


Figure 5.12.3: Raw Summed Score Distribution – Combined

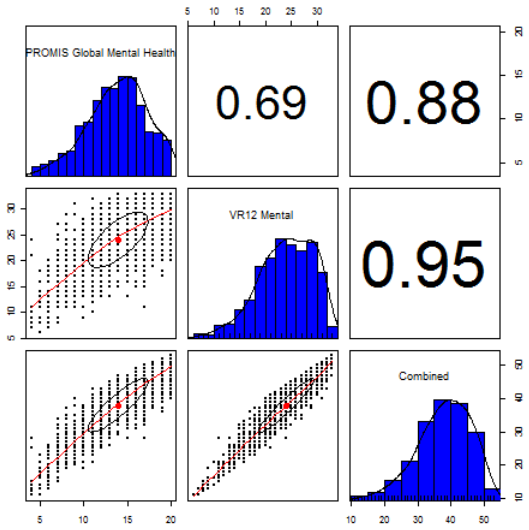


Figure 5.12.4: Scatter Plot Matrix of Raw Summed Scores

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 Global Mental Health, Cronbach's alpha internal consistency reliability estimate was 0.809 and adjusted (corrected for overlap) item-total correlations ranged from 0.477 to 0.704. For VR12 Mental, alpha was 0.802 and adjusted item-total correlations ranged from 0.441 to 0.689. For the 10 items, alpha was 0.872 and adjusted item-total correlations ranged from 0.441 to 0.679.

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 Global Mental Health	4	0.809	0.477	0.630	0.704
VR12 Mental	6	0.802	0.441	0.568	0.689
Combined	10	0.872	0.441	0.598	0.679

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 Global Mental Health, the fit statistics were as follows: CFI = 0.998, TLI = 0.994, and RMSEA = 0.067. For VR12 Mental, CFI = 0.857, TLI = 0.762, and RMSEA = 0.297. For the 10 items, CFI = 0.844, TLI = 0.799, and RMSEA = 0.213. 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 Global Mental Health	4	2025	0.998	0.994	0.067
VR12 Mental	6	2025	0.857	0.762	0.297
Combined	10	2025	0.844	0.799	0.213

5.12.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 10 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 10 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 4 PROMIS Global Mental Health items served as anchor items to transform the item parameter estimates for the VR12 Mental items onto the PROMIS Global Mental Health 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 Global Mental Health items were constrained to their final bank values, while the VR12 Mental items were calibrated, under the constraints imposed by the anchor items.

Table 5.12.3: IRT Linking Constants

	A	B
Mean/Mean	0.738	-0.301
Mean/Sigma	0.958	-0.166
Haebara	0.921	-0.176
Stocking-Lord	0.904	-0.190

The item parameter estimates for the VR12 Mental items were linked to the PROMIS Global Mental Health metric using the transformation constants shown in Table 5.12.3. The VR12 Mental item parameter estimates from the fixed- parameter calibration are considered already on the PROMIS Global Mental Health metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for VR12 Mental 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.

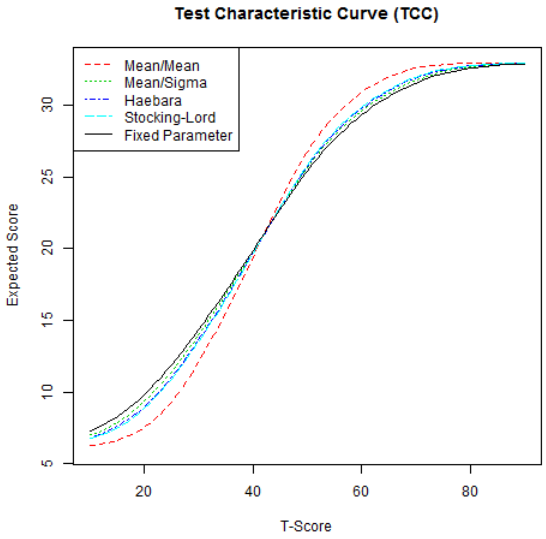


Figure 5.12.5: Test Characteristic Curves (TCC) from Different Linking Methods

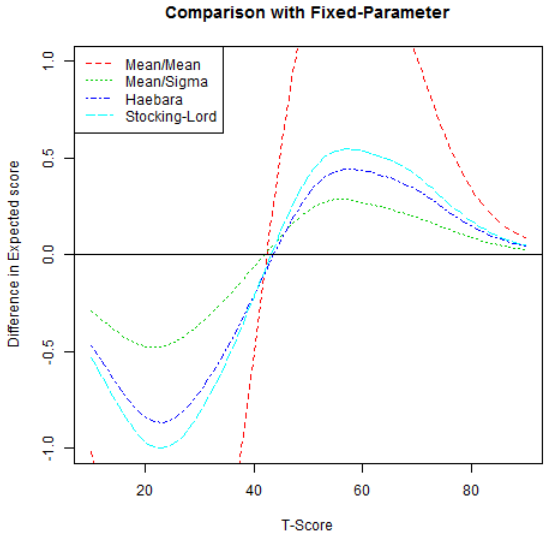


Figure 5.12.6: Difference in Test Characteristic Curves (TCC)

Table 5.12.4 shows the fixed-parameter calibration item parameter estimates for VR12 Mental. The marginal reliability estimate for VR12 Mental based on the item parameter estimates was 0.795. The marginal reliability estimates for PROMIS Global Mental Health and the combined set were 0.86 and 0.909, respectively. The slope parameter estimates for VR12 Mental ranged from 1.19 to 1.9 with a mean of 1.59. The slope parameter estimates for PROMIS Global Mental Health ranged from 1.82 to 3.53 with a mean of 2.63. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.12.7 displays the scale information functions for PROMIS Global Mental Health, VR12 Mental, and the combined set of 10. 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

a	cb1	cb2	cb3	cb4	cb5	NCAT
1.739	-2.405	-1.282	-0.355	0.385	1.876	6
1.689	-2.943	-1.566	-0.745	0.000	1.601	6
1.191	-3.189	-1.971	-1.191	-0.368	1.211	6
1.419	-3.108	-2.021	-0.856	0.112		5
1.899	-2.541	-1.659	-0.788	0.079		5
1.574	-3.159	-2.070	-1.004	-0.097		5

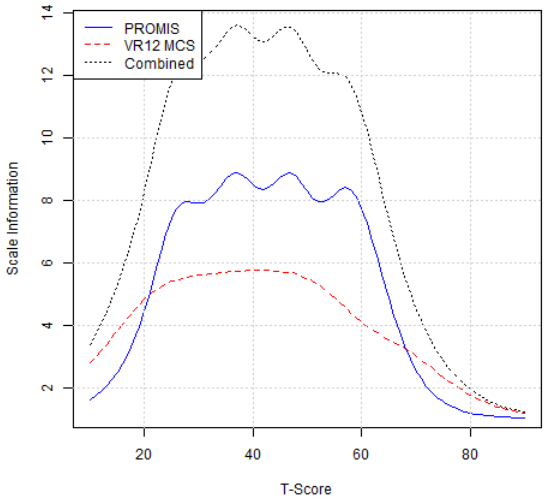


Figure 5.12.7: Comparison of Scale Information Functions

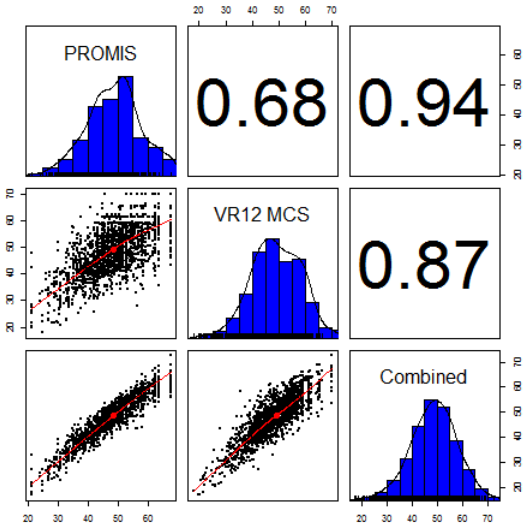


Figure 5.12.8: Comparison of IRT Scaled Scores

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 VR12 Mental to a scaled score on PROMIS Global Mental Health can be useful. Based on the VR12 Mental 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 VR12 Mental to T-score values linked to the PROMIS Global Mental Health 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 VR12 Mental to a corresponding scaled score on PROMIS Global Mental Health by identifying scores on PROMIS Global Mental Health that have the same percentile ranks as scores on VR12 Mental. 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 VR12 Mental to PROMIS Global Mental Health. 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).

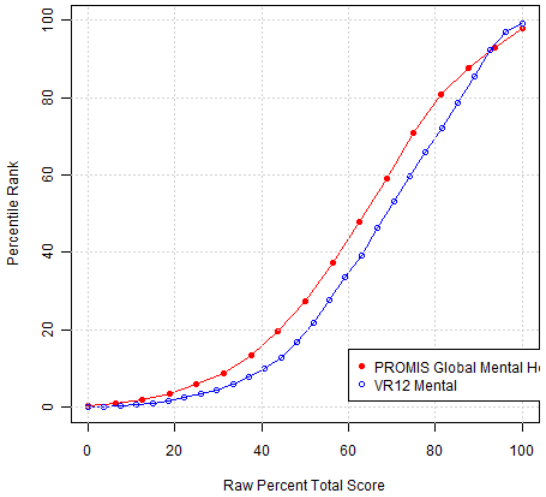


Figure 5.12.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

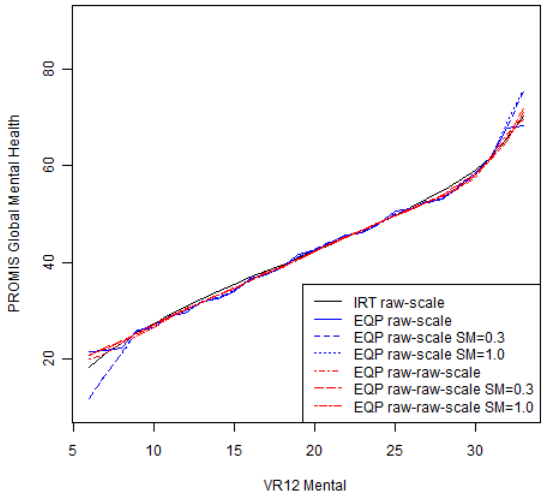


Figure 5.12.10: Equipercentile Linking Functions

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 Global Mental Health T-scores and VR12 Mental 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 (0.675), followed by EQP raw-raw-scale SM=1.0 (0.662). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (7.461), followed by EQP raw-raw-scale SM=0.0 (7.52).

Table 5.12.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.675	-0.260	7.458	7.461
IRT raw-scale	0.661	-0.259	7.529	7.532
EQP raw-scale SM=0.0	0.661	-0.031	7.583	7.581
EQP raw-scale SM=0.3	0.659	-0.066	7.719	7.718
EQP raw-scale SM=1.0	0.658	-0.070	7.762	7.761
EQP raw-raw-scale SM=0.0	0.661	0.140	7.521	7.520
EQP raw-raw-scale SM=0.3	0.661	0.104	7.553	7.551
EQP raw-raw-scale SM=1.0	0.662	0.108	7.525	7.524

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=2017) 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 Global Mental Health 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, 0.85. That is, the difference between the mean PROMIS Global Mental Health T-score and the mean equated VR12 Mental T-score based on a similar sample of 75 cases is expected to be around ±1.7 (i.e., 2 × 0.85).

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	-0.253	1.470	-0.257	1.030	-0.242	0.850
IRT raw-scale	-0.258	1.490	-0.246	1.043	-0.263	0.854
EQP raw-scale SM=0.0	-0.036	1.514	-0.016	1.056	-0.021	0.866
EQP raw-scale SM=0.3	-0.047	1.533	-0.076	1.080	-0.063	0.882
EQP raw-scale SM=1.0	-0.067	1.529	-0.091	1.079	-0.089	0.880
EQP raw-raw-scale SM=0.0	0.140	1.495	0.150	1.052	0.148	0.857
EQP raw-raw-scale SM=0.3	0.111	1.501	0.104	1.061	0.104	0.856
EQP raw-raw-scale SM=1.0	0.119	1.499	0.110	1.052	0.101	0.850

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 Global Health - Mental component and VR-12 – Mental Component (Algorithmic Scores)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Global Health Mental, namely the PROMIS Global Mental Health item bank (4 items) and VR-12 Mental Component Score (MCS; 6 items forming one algorithmic score). Both instruments were scaled such that higher scores represent higher levels of Global Mental Health. Seven participants had 1 or more missing responses, leaving a linking sample of N=2018. We created raw summed scores for the PROMIS Global Mental Health and a single algorithmic score for the VR-12 MCS. Summing of item scores assumes that all items have positive correlations with the total as examined in the section on Classical Item Analysis. Because we linked from VR-12 algorithmic score, we could not apply IRT-based linking (which takes advantage of the pattern of item responses). Therefore, we completed equipercentile linking only.

.

5.13.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 20 for PROMIS Global Mental Health and 33 for VR12 Mental. Figure 5.13.1 shows the distribution for the combined VR-12 MCS and PROMIS scores. Figure 5.13.2 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 Global Mental Health and VR-12 MCS was 0.63. The correlations between the combined score and the measures were 0.77 and 0.98 for PROMIS Global Mental Health and VR-12 MCS, respectively.

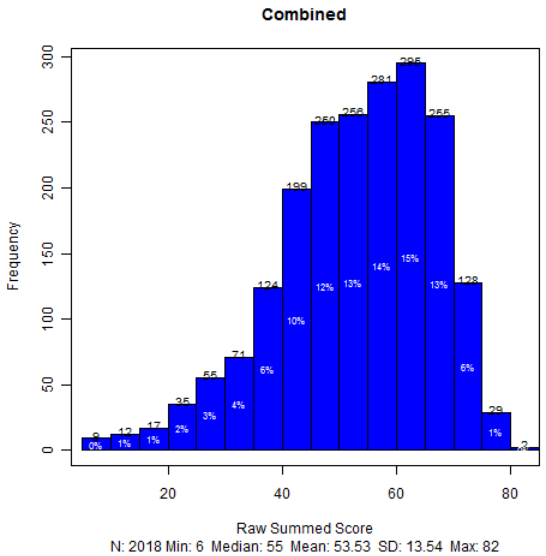


Figure 5.13.1: Raw Summed Score Distribution – Combined PROMIS Global Mental Health and VR-12 Mental (algorithmic scores)

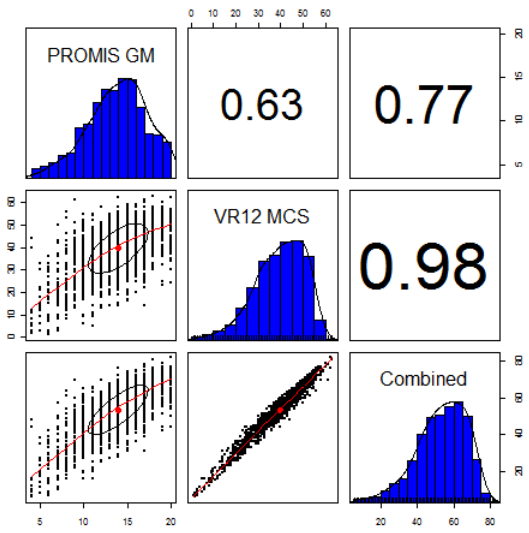


Figure 5.13.2: Scatter Plot Matrix of Raw Summed Scores

5.13.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined scores. Table 5.13.1 summarizes the results. For PROMIS Global Mental Health, the Cronbach's alpha internal consistency reliability estimate (standardized) was 0.81 and adjusted (corrected for overlap) item-total correlations ranged from 0.48 to 0.70. For the 6 items combined, alpha was 0.84 and adjusted item-total correlations ranged from 0.50 to 0.63.

Table 5.13.1: Classical Item Analysis

Instruments	Items	Item-Total Correlations			AIC	Alpha	Omega-h
		Min.	Mean	Max.			
PROMIS Mental Health	4	0.48	0.63	0.70	0.52	0.81	0.79
PROMIS & VR12 MCS (Alg.)	5	0.50	0.58	0.63	0.51	0.84	0.74

Note. Alpha is standardized. AIC = average inter-item correlation. Omega-h = omega hierarchical

5.13.3. Dimensionality of the measures

To assess the relative dimensionality of the measures, we estimated the proportion of total variance attributable to a general factor (ωh; McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005) using the psych package (Revelle, 2013) in R (R Core Development Team, 2011). This method estimates ωh from the general factor loadings derived from an exploratory factor analysis and a Schmid–Leiman transformation (Schmid & Leiman, 1957).The estimate of general factor saturation (ωh) for the combined measure was reasonably high: 0.74 (PROMIS and VR-12). (See Table 5.13.1). This value suggests the presence of a fairly large general factor for each instrument pair (Reise, Scheines, Widaman, & Haviland, 2012).

5.13.4. Equipercentile Linking

We mapped each raw summed score point on VR-12 MCS to a corresponding scaled score on PROMIS Global Mental Health by identifying scores on PROMIS Global Mental Health that have the same percentile ranks as scores on VR-12 MCS. 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 5.13.3 displays the cumulative distribution functions of

the measures. Figure 5.13.4 shows the equipercentile linking functions based on raw summed scores, from VR-12 MCS to PROMIS Global Mental Health. 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. Three separate equipercentile equivalents were computed and compared: one is equipercentile without post smoothing ("Equipercentile Scale Score Equiv- alents") 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). Appendix Table 37 shows the recommended equipercentile crosswalk table.

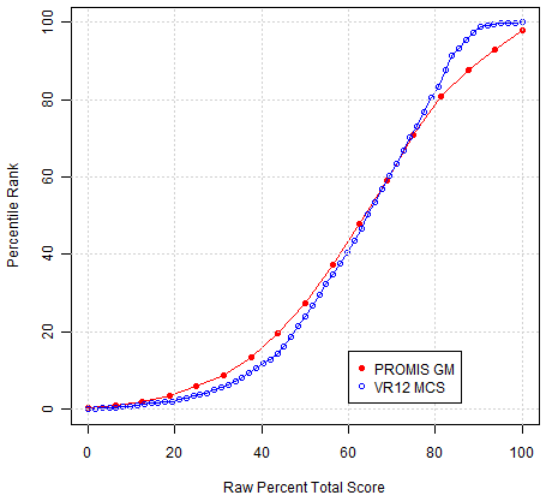


Figure 5.13.3: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

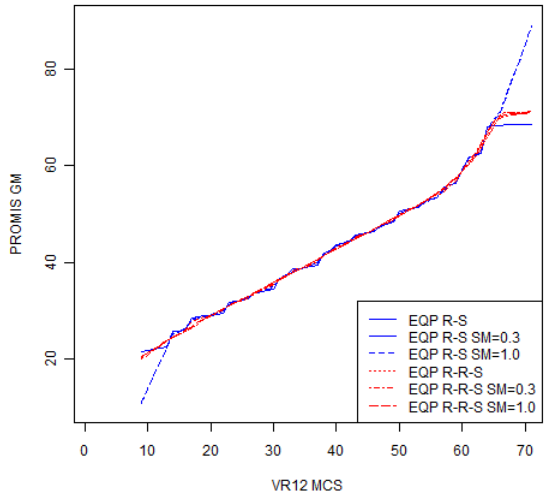


Figure 5.13.4: Equipercentile Linking Functions

Note. R-S = VR-12 "Raw" to PROMIS Scale; R-R-S = VR-12 "Raw" to PROMIS Raw to PROMIS Scale.

5.13.5. 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. As displayed in Figure 5.13.4, 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 5.13.2 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Global Mental Health T-scores and VR-12 MCS scores linked to the T-score metric through different methods. With respect to the correlation between observed and linked T-scores, Equipercentile (Raw-Scale, no smoothing) produced the best result (0.606), though results were similar by method. Similar results were also found in terms of the standard deviation of differences and root mean squared difference (RMSD).

Table 5.13.2: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
EQP raw-scale SM=0.0	0.606	0.08	8.092	8.091
EQP raw-scale SM=0.3	0.599	0.063	8.268	8.267
EQP raw-scale SM=1.0	0.599	0.075	8.279	8.277
EQP raw-raw-scale SM=0.0	0.603	0.133	8.154	8.153
EQP raw-raw-scale SM=0.3	0.603	0.111	8.155	8.153
EQP raw-raw-scale SM=1.0	0.604	0.116	8.135	8.134

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=2017) over a large number of replications (i.e., 10,000).

Table 5.13.3 summarizes the 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 Global Mental Health T-scores was computed. Then the standard deviation of the means was computed over replications as the empirical standard error. As the sample size increased (from 25 to 150), the empirical standard error decreased steadily. At a sample size of 75, EQP raw- raw-scale SM=0.0 produced a standard error of 0.92. This can interpreted in the following way: the difference between the mean PROMIS Global Mental Health T-score and the mean equated VR-12 MCS T-score based on a similar sample of 75 cases is expected to be around ±1.84 (i.e., 2 × 0.92).

Table 5.13.3: Comparison of Resampling Results (Standard Deviations)

Methods	(N=25)	(N=50)	(N=75)	(N=100)	(N=125)	(N=150)
EQP raw-scale SM=0.0	1.62	1.128	0.921	0.797	0.697	0.63
EQP raw-scale SM=0.3	1.654	1.149	0.927	0.789	0.715	0.65
EQP raw-scale SM=1.0	1.644	1.167	0.933	0.817	0.715	0.654
EQP raw-raw-scale SM=0.0	1.618	1.146	0.915	0.795	0.705	0.64
EQP raw-raw-scale SM=0.3	1.612	1.143	0.926	0.794	0.704	0.639
EQP raw-raw-scale SM=1.0	1.602	1.13	0.928	0.798	0.704	0.644

Examining a number of linking studies in the current project revealed that the equipercentile linking methods produced highly comparable results for most scores. Some noticeable discrepancies were observed in some extreme score levels where data were sparse.

5.14. PROMIS Global Health-Physical and VR-12-Physical

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Global Health – Physical component, namely the PROMIS Global Physical Health item bank (4 items) and VR12 Physical (7 items). Both instruments were scaled such that higher scores represent higher levels of Global Health - Physical. We excluded 1 participant because of missing responses, leaving a final sample of N=2020. 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 20 for PROMIS Global Physical Health and 32 for VR12 Physical. 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 Global Physical Health and VR12 Physical was 0.8. The disattenuated (corrected for unreliabilities) correlation between PROMIS Global Physical Health and VR12 Physical was 1. The correlations between the combined score and the measures were 0.92 and 0.97 for PROMIS Global Physical Health and VR12 Physical, respectively.

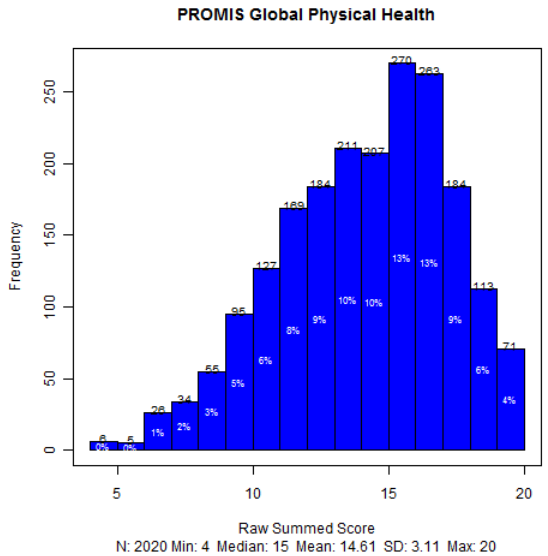


Figure 5.14.1: Raw Summed Score Distribution - PROMIS Global Health – Physical component

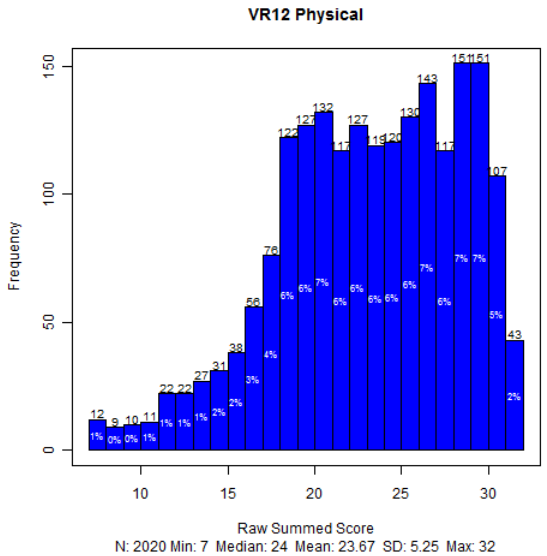


Figure 5.14.2: Raw Summed Score Distribution – VR-12 – Physical component

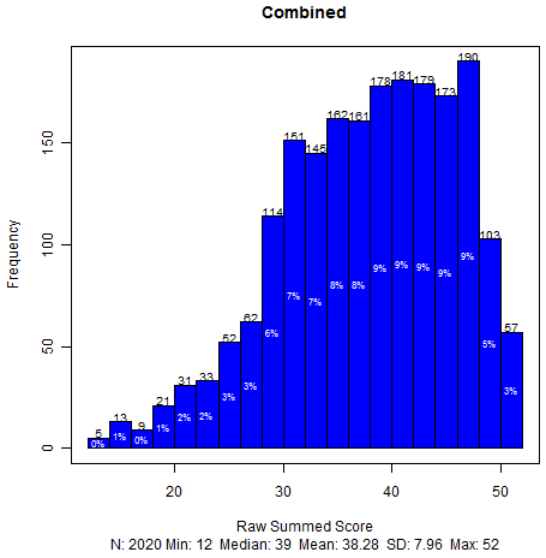


Figure 5.14.3: Raw Summed Score Distribution – Combined

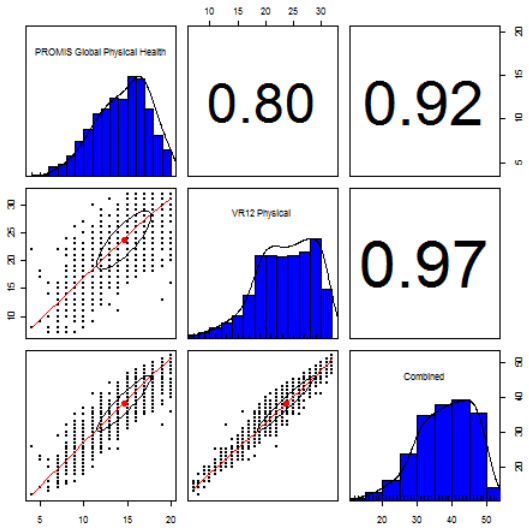


Figure 5.14.4: Scatter Plot Matrix of Raw Summed Scores

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 Global Physical Health, Cronbach's alpha internal consistency reliability estimate was 0.776 and adjusted (corrected for overlap) item-total correlations ranged from 0.551 to 0.603. For VR12 Physical, alpha was 0.83 and adjusted item-total correlations ranged from 0.49 to 0.678. For the 11 items, alpha was 0.892 and adjusted item-total correlations ranged from 0.545 to 0.696.

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 Global Physical Health	4	0.776	0.551	0.580	0.603
VR12 Physical	7	0.830	0.490	0.594	0.678
Combined	11	0.892	0.545	0.625	0.696

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 Global Physical Health, the fit statistics were as follows: CFI = 0.997, TLI = 0.991, and RMSEA = 0.059. For VR12 Physical, CFI = 0.918, TLI = 0.878, and RMSEA = 0.203. For the 11 items, CFI

= 0.888, TLI = 0.86, and RMSEA = 0.19. 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 Global Physical Health	4	2025	0.997	0.991	0.059
VR12 Physical	7	2025	0.918	0.878	0.203
Combined	11	2025	0.888	0.860	0.190

5.14.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 11 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 11 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 4 PROMIS Global Physical Health items served as anchor items to transform the item parameter estimates for the VR12 Physical items onto the PROMIS Global Physical Health. 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 Global Physical Health items were constrained to their final bank values, while the VR12 Physical items were calibrated, under the constraints imposed by the anchor items.

Table 5.14.3: IRT Linking Constants

	A	B
Mean/Mean	0.877	-0.418
Mean/Sigma	1.027	-0.283
Haebara	0.998	-0.307
Stocking-Lord	0.985	-0.318

The item parameter estimates for the VR12 Physical items were linked to the PROMIS Global Physical Health metric using the transformation constants shown in Table 5.14.3. The VR12 Physical item parameter estimates from the fixed- parameter calibration are considered already on the PROMIS Global Physical Health metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for VR12 Physical 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.

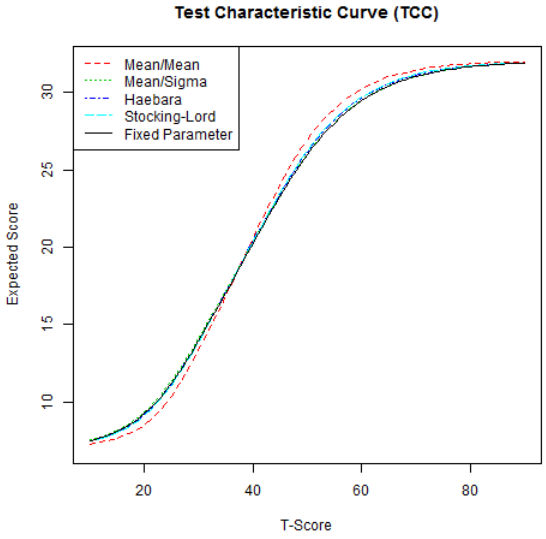


Figure 5.14.5: Test Characteristic Curves (TCC) from Different Linking Methods

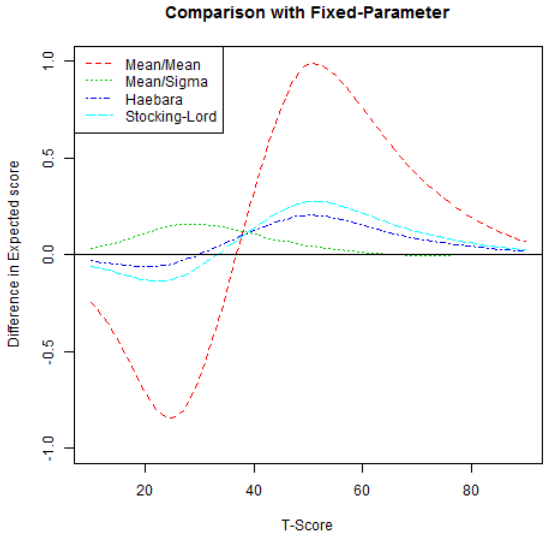


Figure 5.14.6: Difference in Test Characteristic Curves (TCC)

Table 5.14.4 shows the fixed-parameter calibration item parameter estimates for VR12 Physical. The marginal reliability estimate for VR12 Physical based on the item parameter estimates was 0.829. The marginal reliability estimates for PROMIS Global Physical Health and the combined set were 0.792 and 0.895, respectively. The slope parameter estimates for VR12 Physical ranged from 1.31 to 2.51 with a mean of 1.9. The slope parameter estimates for PROMIS Global Physical Health ranged from 1.68 to 2.88 with a mean of 2.15. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.14.7 displays the scale information functions for PROMIS Global Physical Health, VR12 Physical, and the combined set of 11. 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

a	cb1	cb2	cb3	cb4	cb5	NCAT
1.811	-1.619	-0.450				3
1.755	-1.656	-0.215				3
2.002	-2.535	-1.887	-1.020	0.009		5
2.511	-2.485	-1.844	-1.130	-0.260		5
1.924	-2.367	-1.621	-0.867	0.234		5
1.310	-2.968	-1.643	-0.546	0.350	2.203	6
2.009	-2.645	-1.444	-0.250	1.150		5

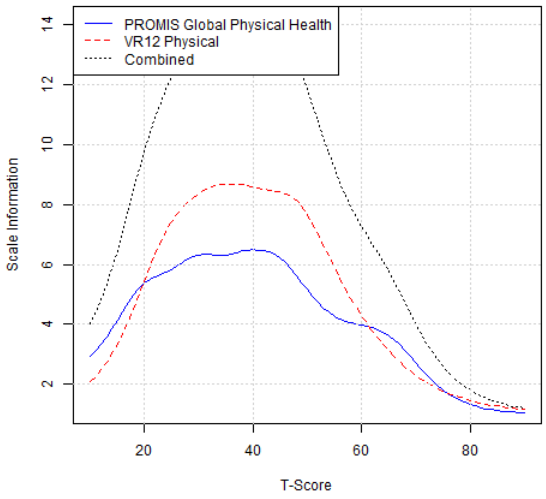


Figure 5.14.7: Comparison of Scale Information Functions

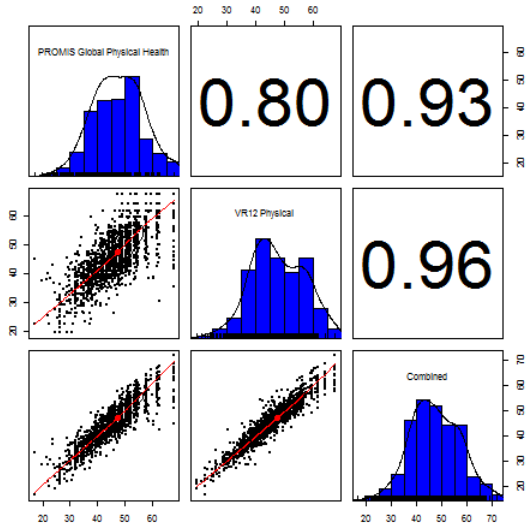


Figure 5.14.8: Comparison of IRT Scaled Scores

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 VR12 Physical to a scaled score on PROMIS Global Physical Health can be useful. Based on the VR12 Physical item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 38 can be used to map simple raw summed scores from VR12 Physical to T-score values linked to the PROMIS Global Physical Health 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 VR12 Physical to a corresponding scaled score on PROMIS Global Physical Health by identifying scores on PROMIS Global Physical Health that have the same percentile ranks as scores on VR12 Physical. 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 VR12 Physical to PROMIS Global Physical Health.

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 39 and Appendix Table 40 show the equipercentile crosswalk tables. The result shown in Appendix Table 39 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 40 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).

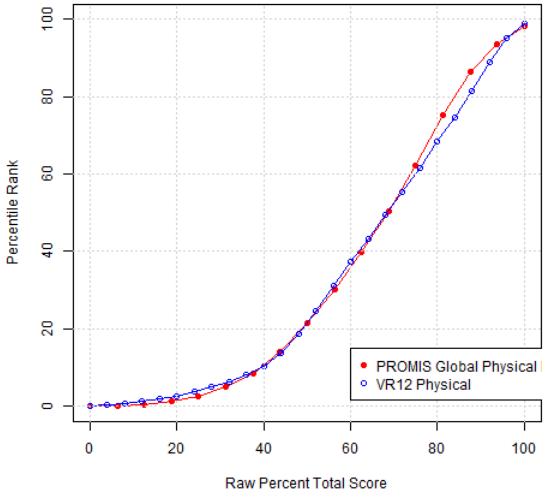


Figure 5.14.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

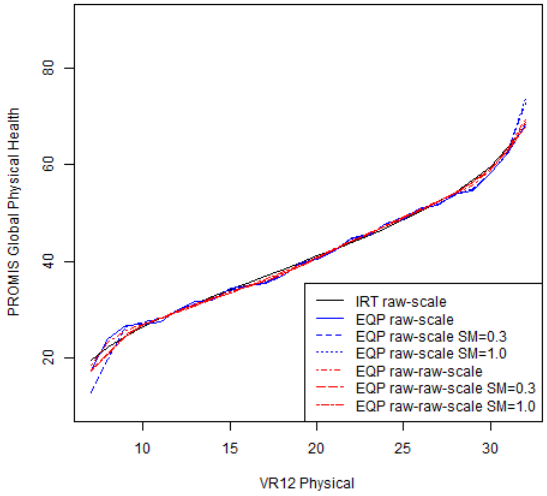


Figure 5.14.10: Equipercentile Linking Functions

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 Global Physical Health T-scores and VR12 Physical 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 (0.801), followed by IRT raw-scale (0.797). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). EQP raw-scale SM=0.0 yielded smallest RMSD (5.88), followed by IRT raw-scale (5.909).

Table 5.14.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.801	-0.302	5.912	5.919
IRT raw-scale	0.797	-0.291	5.903	5.909
EQP raw-scale SM=0.0	0.793	0.111	5.880	5.880
EQP raw-scale SM=0.3	0.790	-0.036	6.030	6.029
EQP raw-scale SM=1.0	0.789	-0.085	6.062	6.061
EQP raw-raw-scale SM=0.0	0.796	-0.161	5.923	5.924
EQP raw-raw-scale SM=0.3	0.796	-0.127	5.929	5.929
EQP raw-raw-scale SM=1.0	0.797	-0.114	5.918	5.918

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=2020) 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 Global Physical Health 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, 0.665. That is, the difference between the mean PROMIS Global Physical Health T-score and the mean equated VR12 Physical T-score based on a similar sample of 75 cases is expected to be around ±1.33 (i.e., 2 × 0.665).

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	-0.308	1.185	-0.312	0.825	-0.300	0.672
IRT raw-scale	-0.289	1.170	-0.304	0.830	-0.291	0.665
EQP raw-scale SM=0.0	0.122	1.154	0.104	0.810	0.112	0.669
EQP raw-scale SM=0.3	-0.054	1.202	-0.051	0.836	-0.028	0.683
EQP raw-scale SM=1.0	-0.088	1.225	-0.076	0.851	-0.089	0.681
EQP raw-raw-scale SM=0.0	-0.168	1.182	-0.164	0.839	-0.166	0.668
EQP raw-raw-scale SM=0.3	-0.131	1.170	-0.115	0.825	-0.135	0.673
EQP raw-raw-scale SM=1.0	-0.118	1.178	-0.114	0.825	-0.128	0.672

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

5.15. PROMIS Global Health - Physical Component and VR-12 – Physical Component (Algorithmic Scores)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Global Health Physical, namely the PROMIS Global Physical Health item bank (4 items) and VR-12 Physical Component Score (PCS; 7 items forming one algorithmic score). Both instruments were scaled such that higher scores represent higher levels of Global Physical Health. Two participants had 1 or more missing responses, leaving a linking sample of N=2023. We created raw summed scores for the PROMIS Global Physical Health and a single algorithmic score for the VR-12 PCS. Summing of item scores assumes that all items have positive correlations with the total as examined in the section on Classical Item Analysis. Because we linked from VR-12 algorithmic score, we could not apply IRT-based linking (which takes advantage of the pattern of item responses). Therefore, we completed equipercentile linking only.

5.15.1. Raw Summed Score Distribution

The maximum possible raw summed score was 20 for PROMIS Global Physical Health. The VR-12 algorithmic scores used for linking were rounded to the nearest integer and ranged from 10 to 65 in our linking sample. Figure 5.15.1 shows the distribution for the combined VR-12 PCS and PROMIS scores. Figure 5.15.2 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 Global Physical Health and VR-12 PCS was 0.69. The correlations between the combined score and the measures were 0.81 and 0.98 for PROMIS Global Physical Health and VR-12 PCS, respectively.

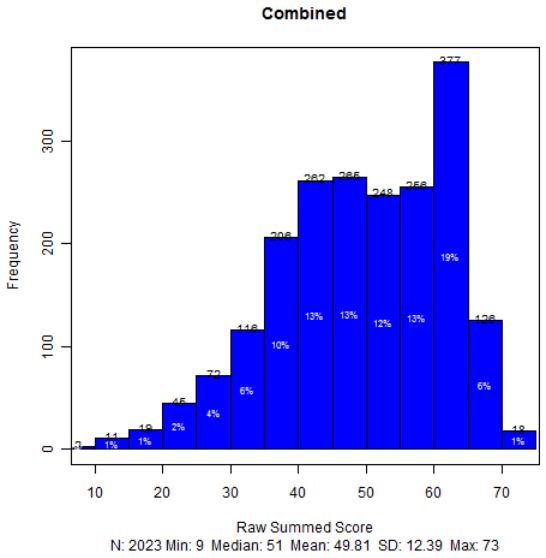


Figure 5.15.1: Raw Summed Score Distribution – Combined PROMIS Global Physical Health and VR-12 Physical (algorithmic scores)

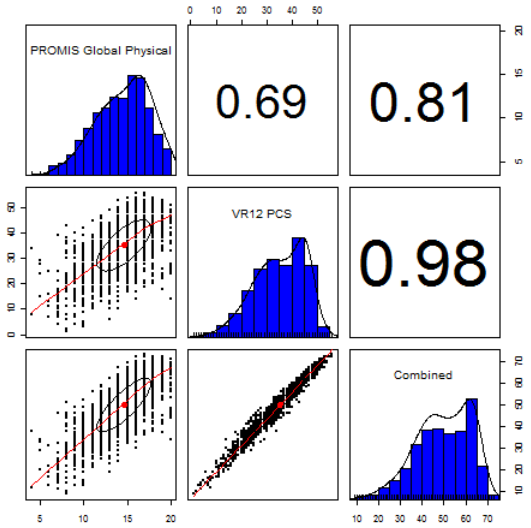


Figure 5.15.2: Scatter Plot Matrix of Raw Summed Scores

5.15.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined scores. Table 5.15.1 summarizes the results. For PROMIS Global Physical Health, the Cronbach's alpha internal consistency reliability estimate (standardized) was 0.83 and adjusted (corrected for overlap) item-total correlations ranged from 0.55 to 0.60. For the 5 items combined, alpha was 0.84 and adjusted item-total correlations ranged from 0.47 to 0.60.

Table 5.15.1: Classical Item Analysis

		Item-Total Correlations					
	Items				AIC	Alpha	Omega-h
Instruments		Min.	Mean	Max.			
PROMIS Physical Health	4	0.55	0.58	0.60	0.47	0.78	0.74
PROMIS & VR12 PCS (Alg.)	5	0.47	0.60	0.70	0.49	0.83	0.74

Note. Alpha is standardized. AIC = average inter-item correlation. Omega-h = omega hierarchical

5.15.3. Dimensionality of the measures

To assess the relative dimensionality of the measures, we estimated the proportion of total variance attributable to a general factor (ω_h; McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005) using the **psych** package (Revelle, 2013) in **R** (R Core

Development Team, 2011). This method estimates ω_h from the general factor loadings derived from an exploratory factor analysis and a Schmid–Leiman transformation (Schmid & Leiman, 1957).The estimate of general factor saturation (ω_h) for the combined measure was reasonably high: 0.74 (PROMIS and VR-12). (See Table 5.15.1). This value suggests the presence of a fairly large general factor for each instrument pair (Reise, Scheines, Widaman, & Haviland, 2012).

5.15.4. Equipercentile Linking

We mapped each raw summed score point on VR-12 PCS to a corresponding scaled score on PROMIS Global Physical Health by identifying scores on PROMIS Global Physical Health that have the same percentile ranks as scores on VR-12 PCS. 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 5.15.3 displays the cumulative distribution functions of the measures. Figure 5.15.4 shows the equipercentile linking functions based on raw summed scores, from VR-12 PCS to PROMIS Global Physical Health. 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. Three separate equipercentile equivalents were computed and compared: one is equipercentile without post smoothing (“Equipercentile Scale Score Equiv- alents”) 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). Appendix Table 41 shows the recommended equipercentile crosswalk table.

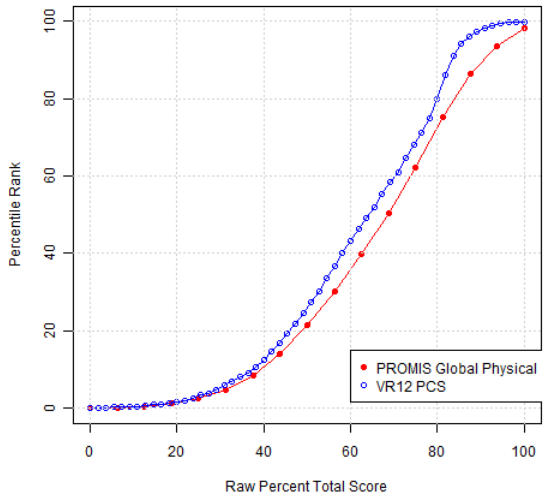


Figure 5.15.3: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

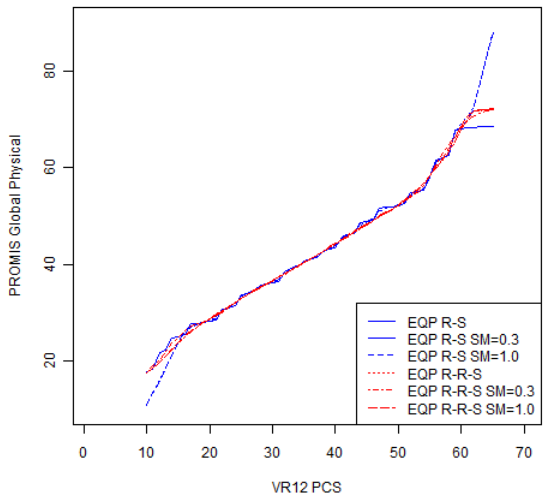


Figure 5.15.4: Equipercentile Linking Functions

Note. R-S = VR-12 "Raw" to PROMIS Scale; R-R-S = VR-12 "Raw" to PROMIS Raw to PROMIS Scale.

5.15.5. 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. As displayed in Figure 5.15.4, 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 5.15.2 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Global Physical Health T-scores and VR-12 PCS scores linked to the T-score metric through different methods. With respect to the correlation between observed and linked T-scores, Equipercentile (Raw-Raw-Scale, high smoothing) produced the best result (0.674), though results in terms of the standard deviation of differences and root mean squared difference (RMSD) suggested other links were slightly more accurate.

Table 5.15.2: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
EQP raw-scale SM=0.0	0.674	-0.208	7.523	7.524
EQP raw-scale SM=0.3	0.665	-0.265	7.724	7.726
EQP raw-scale SM=1.0	0.665	-0.28	7.743	7.746
EQP raw-raw-scale SM=0.0	0.674	-0.175	7.494	7.494
EQP raw-raw-scale SM=0.3	0.675	-0.119	7.486	7.485
EQP raw-raw-scale SM=1.0	0.678	-0.086	7.435	7.433

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=2023) over a large number of replications (i.e., 10,000).

Table 5.15.3 summarizes the 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 Global Physical Health T-scores was computed. Then the standard deviation of the means was computed over replications as the empirical standard error. As the sample size increased (from 25 to 150), the empirical standard error decreased steadily. At a sample size of 75, EQP raw- raw-scale SM=0.0 produced a standard error of 0.86. This can interpreted in the following way: the difference between the mean PROMIS Global Physical Health T-score and the mean equated VR-12 PCS T-score based on a similar sample of 75 cases is expected to be around ±1.72 (i.e., 2 × 0.86).

Table 5.15.3: Comparison of Resampling Results (Standard Deviations)

Methods	(N=25)	(N=50)	(N=75)	(N=100)	(N=125)	(N=150)
EQP raw-scale SM=0.0	1.497	1.049	0.841	0.728	0.65	0.588
EQP raw-scale SM=0.3	1.543	1.067	0.888	0.746	0.67	0.615
EQP raw-scale SM=1.0	1.529	1.091	0.861	0.758	0.672	0.605
EQP raw-raw-scale SM=0.0	1.499	1.041	0.858	0.729	0.652	0.589
EQP raw-raw-scale SM=0.3	1.504	1.056	0.848	0.722	0.645	0.585
EQP raw-raw-scale SM=1.0	1.484	1.033	0.844	0.715	0.647	0.574

Examining a number of linking studies in the current project revealed that the equipercentile linking methods produced highly comparable results for most scores. Some noticeable discrepancies were observed in some extreme score levels where data were sparse.

5.16. PROMIS Pain Interference and SF-36/BP

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 item bank (41 items) and SF-36/BP (2 items). PROMIS Pain Interference 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. Our sample consisted of 730 participants (N = 694 for participants with complete responses).

5.16.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 205 for PROMIS Pain Interference and 11 for SF-36 BP. 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 Pain Interference and SF-36 BP was 0.84. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pain Interference and SF-36 BP was 0.93. The correlations between the combined score and the measures were 1 and 0.85 for PROMIS Pain Interference and SF-36 BP, respectively.

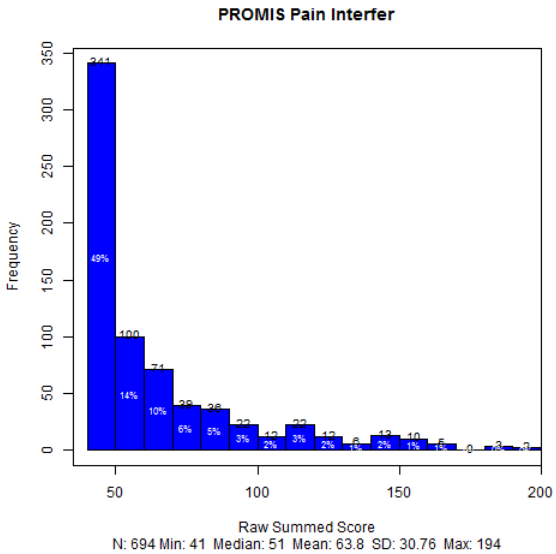


Figure 5.16.1: Raw Summed Score Distribution - PROMIS Pain Interference

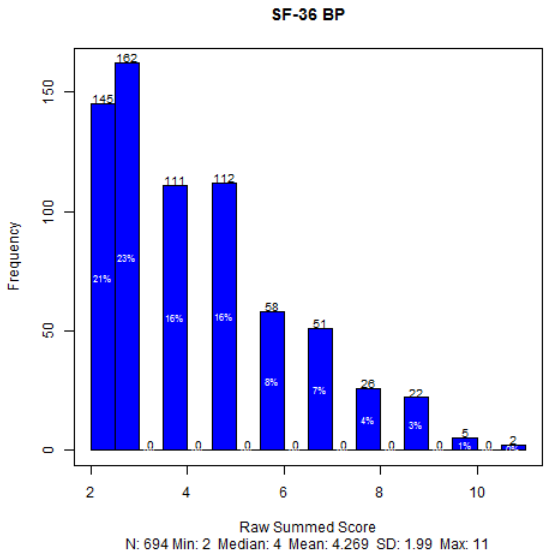


Figure 5.16.2: Raw Summed Score Distribution – SF-36 BP

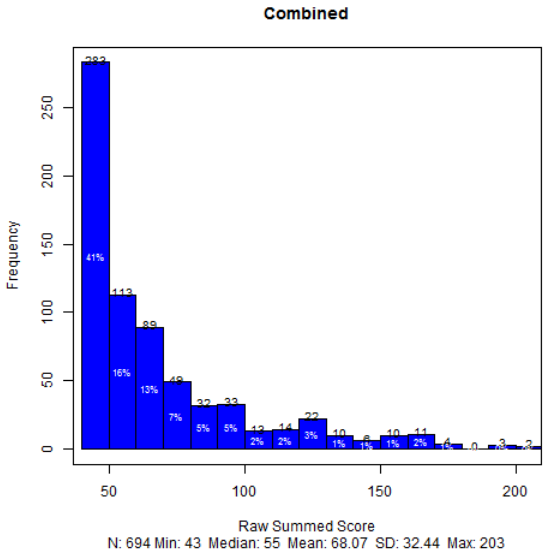


Figure 5.16.3: Raw Summed Score Distribution – Combined

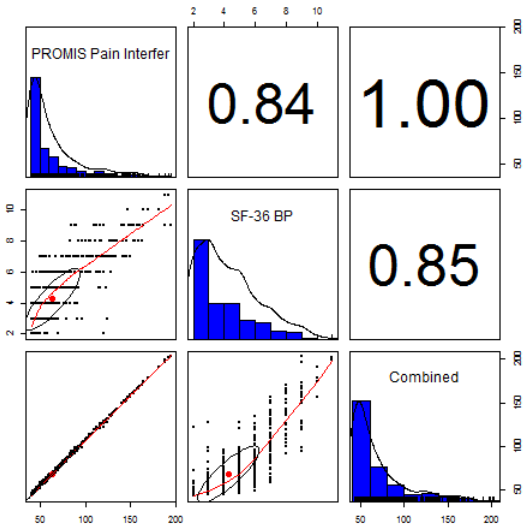


Figure 5.16.4: Scatter Plot Matrix of Raw Summed Scores

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 Pain Interference, Cronbach's alpha internal consistency reliability estimate was 0.986 and adjusted (corrected for overlap) item-total correlations ranged from 0.59 to 0.894. For SF-36 BP, alpha was 0.815 and adjusted item-total correlations ranged from 0.705 to 0.705. For the 43 items, alpha was 0.987 and adjusted item-total correlations ranged from 0.591 to 0.896.

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 Pain Interference	41	0.986	0.590	0.798	0.894
SF-36 BP	2	0.815	0.705	0.705	0.705
Combined	43	0.987	0.591	0.797	0.896

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 Pain Interference and the combined item set. Table 5.16.2 summarizes the model fit statistics.

Table 5.16.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Pain Interference	41	730	0.974	0.972	0.083
Combined	43	730	0.974	0.972	0.081

5.16.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 43 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 43 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 41 PROMIS Pain Interference items served as anchor items to transform the item parameter estimates for the SF-36 BP items onto the PROMIS Pain Interference 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 Pain Interference items were constrained to their final bank values, while the SF-36 BP items were calibrated under the constraints imposed by the anchor items.

Table 5.16.3: IRT Linking Constants

	A	B
Mean/Mean	1.230	0.786
Mean/Sigma	1.230	0.785
Haebara	1.231	0.797
Stocking-Lord	1.227	0.790

The item parameter estimates for the SF-36 BP items were linked to the PROMIS Pain Interference metric using the transformation constants shown in Table 5.16.3. The SF-36 BP item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Pain Interference metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for SF-36 BP 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.

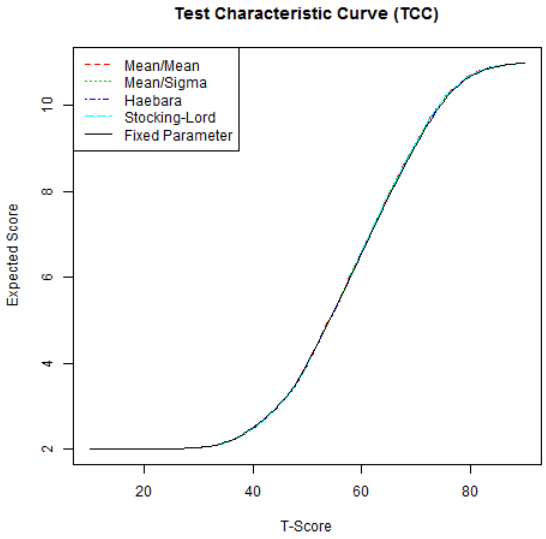


Figure 5.16.5: Test Characteristic Curves (TCC) from Different Linking Methods

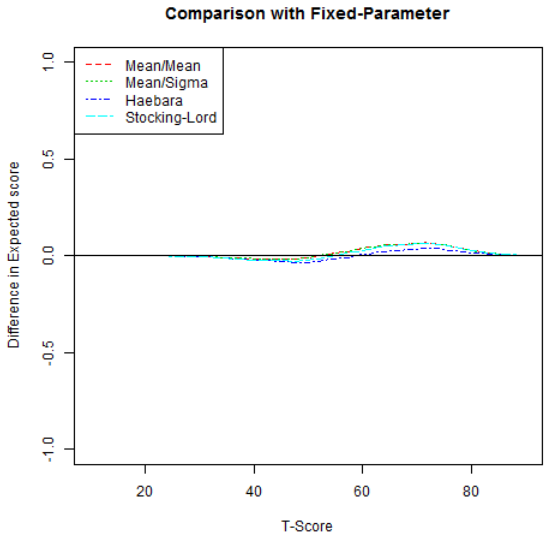


Figure 5.16.6: Difference in Test Characteristic Curves (TCC)

Table 5.16.4 shows the fixed-parameter calibration item parameter estimates for SF-36 BP. The marginal reliability estimate for SF-36 BP based on the item parameter estimates was 0.789. The marginal reliability estimates for PROMIS Pain Interference and the combined set were 0.885 and 0.919, respectively. The slope parameter estimates for SF-36 BP ranged from 2.94 to 4.31 with a mean of 3.63. The slope parameter estimates for PROMIS Pain Interference 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 5.16.7 displays the scale information functions for PROMIS Pain Interference, SF-36 BP, and the combined set of 43. 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 for SF-36 BP

a	cb1	cb2	cb3	cb4	cb5	NCAT
2.942	-0.916	0.035	0.706	1.691	2.689	6
4.309	0.075	0.836	1.427	2.164		5

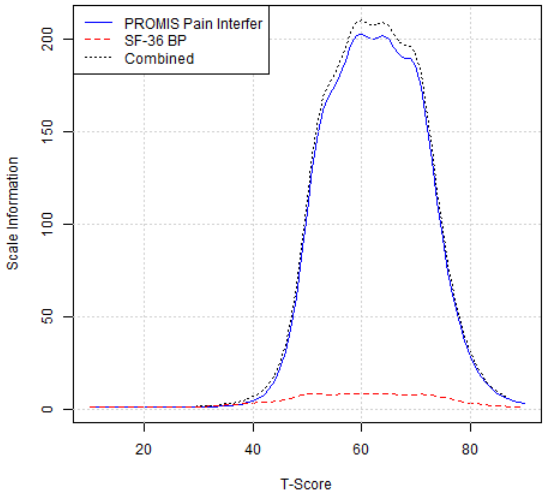


Figure 5.16.7: Comparison of Scale Information Functions

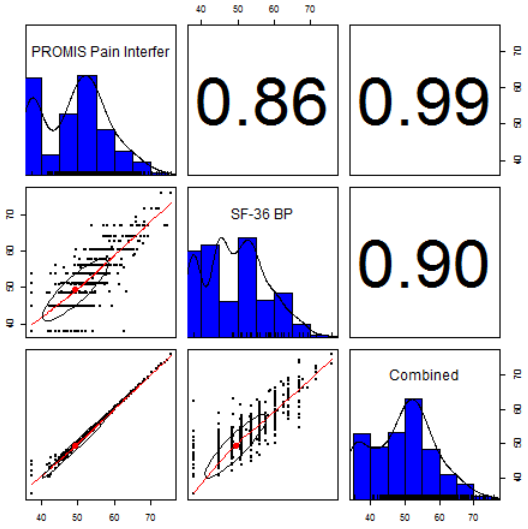


Figure 5.16.8: Comparison of IRT Scaled Scores

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 SF-36 BP to a scaled score on PROMIS Pain Interference can be useful. Based on the SF-36 BP item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 42 can be used to map simple raw summed scores from SF-36 BP to T-score values linked to the PROMIS Pain Interference 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 SF-36 BP to a corresponding scaled score on PROMIS Pain Interference by identifying scores on PROMIS Pain Interference that have the same percentile ranks as scores on SF-36 BP. 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 SF-36 BP to PROMIS Pain Interference. 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 43

and Appendix Table 44 show the equipercentile crosswalk tables. The result shown in Appendix Table 43 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 44 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).

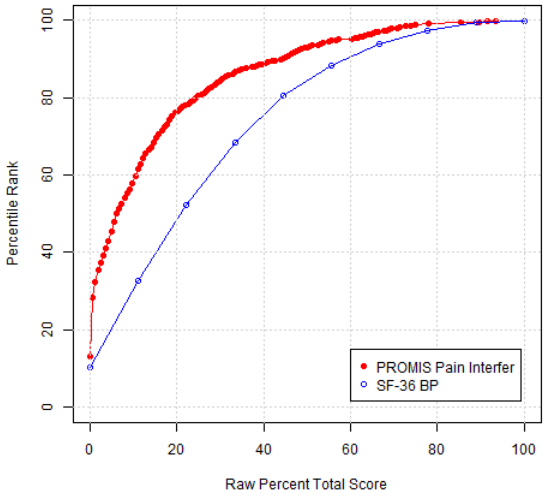


Figure 5.16.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

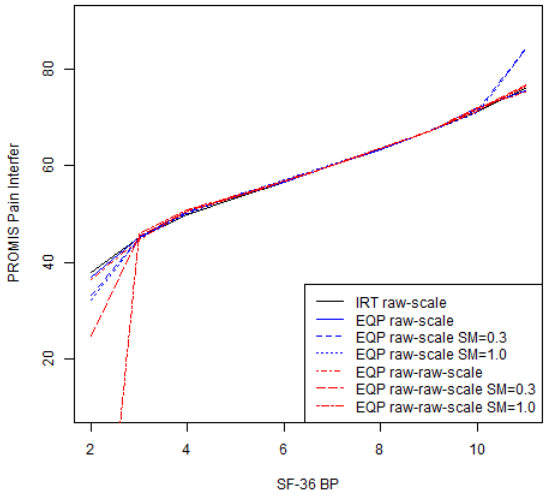


Figure 5.16.10: Equipercentile Linking Functions

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 Pain Interference T-scores and SF-36 BP 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 (0.86), followed by IRT raw-scale (0.852). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (4.668), followed by IRT raw-scale (4.787).

Table 5.16.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.860	-0.165	4.669	4.668
IRT raw-scale	0.852	-0.132	4.789	4.787
EQP raw-scale SM=0.0	0.850	-0.095	4.863	4.861
EQP raw-scale SM=0.3	0.840	0.720	5.449	5.492
EQP raw-scale SM=1.0	0.836	0.958	5.683	5.759
EQP raw-raw-scale SM=0.0	0.850	0.081	4.915	4.912
EQP raw-raw-scale SM=0.3	0.800	2.358	7.750	8.095
EQP raw-raw-scale SM=1.0	0.653	18.668	38.574	42.829

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=694) 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 Pain Interference 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, 0.506. That is, the difference between the mean PROMIS Pain Interference T-score and the mean equated SF-36 BP T-score based on a similar sample of 75 cases is expected to be around ±1.01 (i.e., 2 × 0.506).

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	-0.149	0.923	-0.177	0.629	-0.169	0.506
IRT raw-scale	-0.144	0.931	-0.127	0.654	-0.123	0.522
EQP raw-scale SM=0.0	-0.105	0.962	-0.099	0.657	-0.090	0.530
EQP raw-scale SM=0.3	0.723	1.069	0.707	0.741	0.730	0.602
EQP raw-scale SM=1.0	0.959	1.131	0.966	0.759	0.962	0.615
EQP raw-raw-scale SM=0.0	0.080	0.956	0.080	0.666	0.079	0.538
EQP raw-raw-scale SM=0.3	2.346	1.536	2.352	1.054	2.378	0.847

EQP raw-raw-scale SM=1.0	18.630	7.542	18.647	5.269	18.653	4.251

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

5.17. PROMIS Sleep Disturbance and Neuro-QoL Sleep Disturbance

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Sleep Disturbance, namely the PROMIS Sleep Disturbance item bank (27 items) and Neuro-QoL Sleep Disturbance (8 items). Both instruments were scaled such that higher scores represent higher levels of Sleep Disturbance. We excluded 4 participants because of missing responses, leaving a final sample of N=1012. 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 135 for PROMIS Sleep Disturbance and 40 for Neuro-QoL Sleep Disturbance. Figure 5.17.1 and Figure 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 Sleep Disturbance and Neuro-QoL Sleep Disturbance was 0.81. The disattenuated (corrected for unreliabilities) correlation between PROMIS Sleep Disturbance and Neuro-QoL Sleep Disturbance was 0.88. The correlations between the combined score and the measures were 0.99 and 0.88 for PROMIS Sleep Disturbance and Neuro-QoL Sleep Disturbance, respectively.

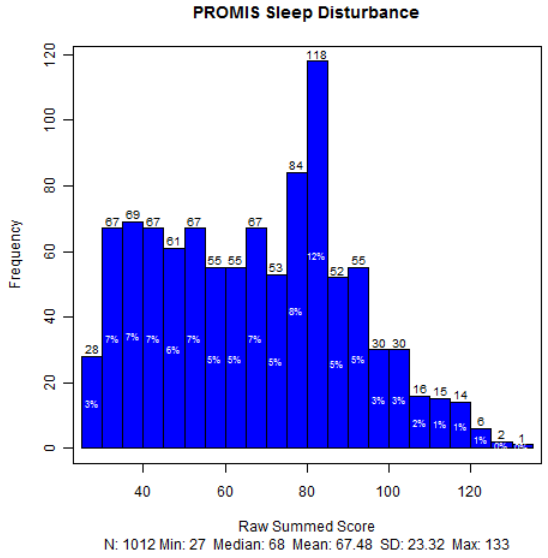


Figure 5.17.1: Raw Summed Score Distribution - PROMIS Sleep Disturbance

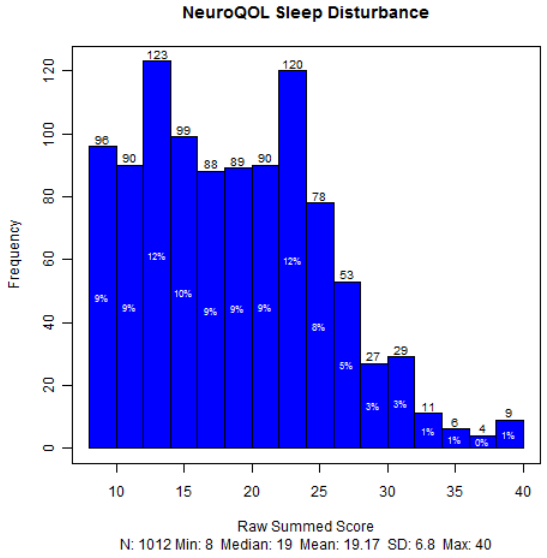


Figure 5.17.2: Raw Summed Score Distribution – Neuro-QOL Sleep Disturbance

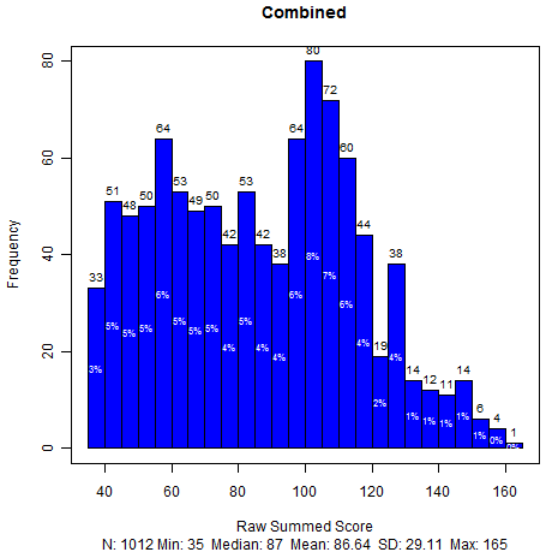


Figure 5.17.3: Raw Summed Score Distribution – Combined

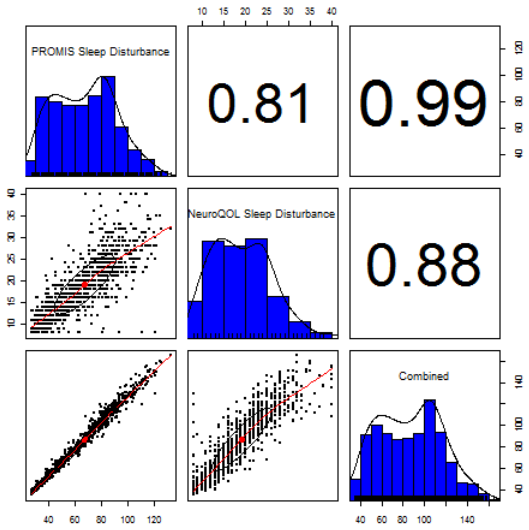


Figure 5.17.4: Scatter Plot Matrix of Raw Summed Scores

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 Sleep Disturbance, Cronbach's alpha internal consistency reliability estimate was 0.965 and adjusted (corrected for overlap) item-total correlations ranged from 0.512 to 0.828. For Neuro-QoL Sleep Disturbance, alpha was 0.877 and adjusted item- total correlations ranged from 0.504 to 0.735. For the 35 items, alpha was 0.969 and adjusted item-total correlations ranged from 0.465 to 0.819.

Table 5.17.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Sleep Disturbance	27	0.965	0.512	0.694	0.828
Neuro-QoL Sleep Disturbance	8	0.877	0.504	0.639	0.735
Combined	35	0.969	0.465	0.676	0.819

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 Sleep Disturbance, the fit statistics were as follows: CFI= 0.872, TLI = 0.861, and RMSEA = 0.158. For Neuro-QoL Sleep Disturbance, CFI = 0.949, TLI = 0.929, and RMSEA =

0.148. For the 35 items, CFI = 0.873, TLI = 0.865, and RMSEA = 0.133. 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 Sleep Disturbance	27	1012	0.872	0.861	0.158
Neuro-QoL Sleep Disturbance	8	1012	0.949	0.929	0.148
Combined	35	1012	0.873	0.865	0.133

5.17.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 35 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 35 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 27 PROMIS Sleep Disturbance items served as anchor items to transform the item parameter estimates for the Neuro-QoL Sleep Disturbance items onto the PROMIS Sleep Disturbance 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 Sleep Disturbance items were con- strained to their final bank values, while the Neuro-QoL Sleep Disturbance items were calibrated, under the constraints imposed by the anchor items.

Table 5.17.3: IRT Linking Constants

	A	B
Mean/Mean	1.072	0.363
Mean/Sigma	1.046	0.374
Haebara	0.995	0.382
Stocking-Lord	1.059	0.343

The item parameter estimates for the Neuro-QoL Sleep Disturbance items were linked to the PROMIS Sleep Disturbance metric using the transformation constants shown in Table 5.17.3. The Neuro-QoL Sleep Disturbance item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Sleep Disturbance metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for Neuro-QoL Sleep Disturbance 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.

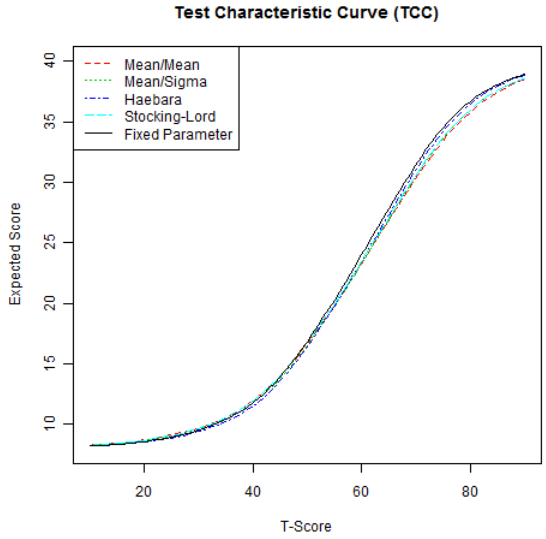


Figure 5.17.5: Test Characteristic Curves (TCC) from Different Linking Methods

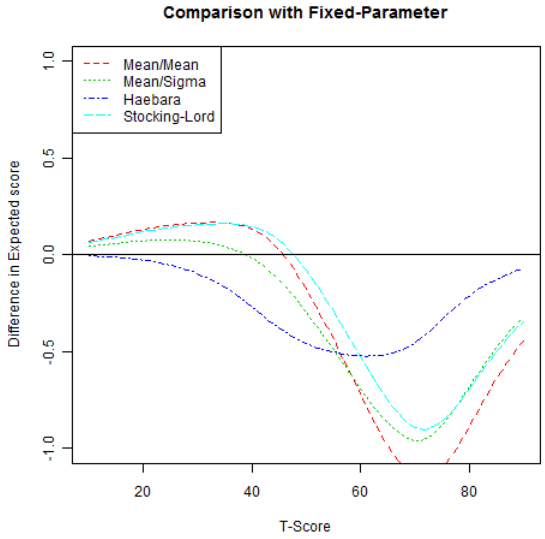


Figure 5.17.6: Difference in Test Characteristic Curves (TCC)

Table 5.17.4 shows the fixed-parameter calibration item parameter estimates Neuro-QoL Sleep Disturbance. The marginal reliability estimate for Neuro-QoL Sleep Disturbance based on the item parameter estimates was 0.857. The marginal reliability estimates for PROMIS Sleep Disturbance and the combined set were 0.964 and 0.97, respectively. The slope parameter estimates for Neuro-QoL Sleep Disturbance ranged from 1.05 to 2.88 with a mean of 1.81. The slope parameter estimates for PROMIS Sleep Disturbance ranged from 1.19 to 3.66 with a mean of 2.15. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.17.7 displays the scale information functions for PROMIS Sleep Disturbance, Neuro-QoL Sleep Disturbance, and the combined set of 35. 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

a	cb1	cb2	cb3	cb4	NCAT
1.051	-1.437	0.077	1.576	2.968	5
1.960	-0.943	0.054	1.180	2.087	5
1.306	-2.080	-0.510	1.073	2.482	5
1.423	-0.048	1.059	1.996	3.051	5
2.879	-0.542	0.336	1.145	1.909	5
1.437	-0.027	0.790	1.864	2.842	5
2.133	0.244	0.885	1.672	2.427	5
2.261	-0.057	0.658	1.554	2.269	5

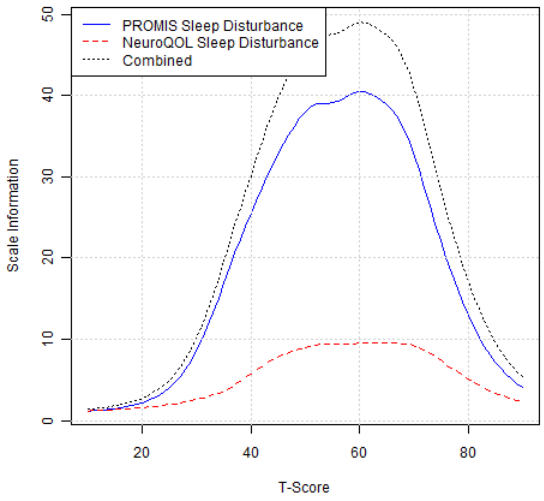


Figure 5.17.7: Comparison of Scale Information Functions

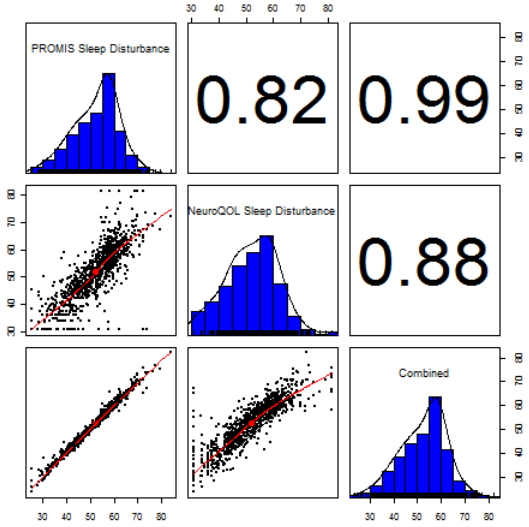


Figure 5.17.8: Comparison of IRT Scaled Scores

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 Sleep Disturbance to a scaled score on PROMIS Sleep Disturbance can be useful. Based on the Neuro-QoL Sleep Disturbance item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 45 can be used to map simple raw summed scores from Neuro-QoL Sleep Disturbance to T-score values linked to the PROMIS Sleep Disturbance 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 Sleep Disturbance to a corresponding scaled score on PROMIS Sleep Disturbance by identifying scores on PROMIS Sleep Disturbance that have the same percentile ranks as scores on Neuro-QoL Sleep Disturbance. 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 Sleep

Disturbance to PROMIS Sleep Disturbance. 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 46 and Appendix Table 47 show the equipercentile crosswalk tables. The result shown in Appendix Table 46 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 47 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).

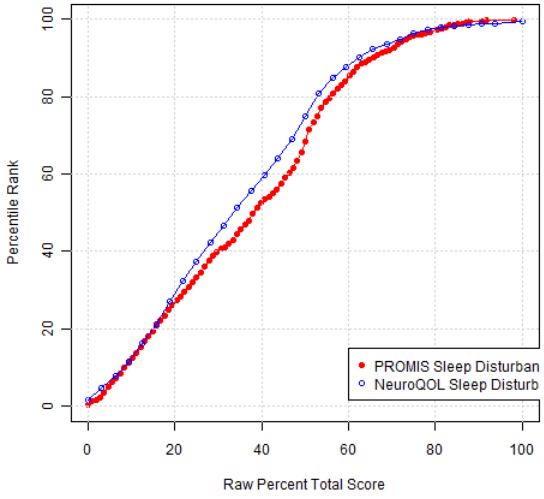


Figure 5.17.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

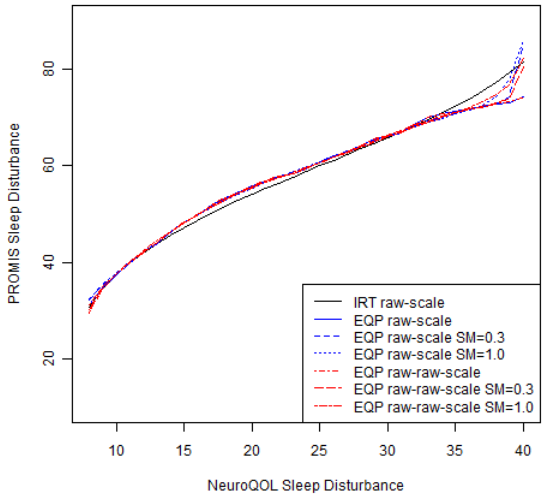


Figure 5.17.10: Equipercentile Linking Functions

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 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 5.17.10, 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 5.17.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Sleep Disturbance T-scores and Neuro-QoL Sleep Disturbance 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 (0.816), followed by EQP raw-scale SM=0.0 (0.804). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (5.942), followed by EQP raw-raw-scale SM=0.0 (6.062).

Table 5.17.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.816	0.367	5.933	5.942
IRT raw-scale	0.795	0.512	6.242	6.260
EQP raw-scale SM=0.0	0.804	-0.202	6.071	6.072
EQP raw-scale SM=0.3	0.798	-0.325	6.217	6.223
EQP raw-scale SM=1.0	0.796	-0.325	6.234	6.239
EQP raw-raw-scale SM=0.0	0.804	-0.191	6.062	6.062
EQP raw-raw-scale SM=0.3	0.799	-0.224	6.186	6.187
EQP raw-raw-scale SM=1.0	0.796	-0.205	6.264	6.264

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1012) over a large number of replications (i.e., 10,000).
Table 5.17.6 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 Sleep Disturbance 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, 0.667. That is, the difference between the mean PROMIS Sleep Disturbance T-score and the mean equated Neuro-QoL Sleep Disturbance T-score based on a similar sample of 75 cases is expected to be around ±1.33 (i.e., 2 × 0.667).

Table 5.17.6: 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	0.372	1.174	0.365	0.813	0.363	0.667
IRT raw-scale	0.515	1.230	0.507	0.861	0.515	0.695
EQP raw-scale SM=0.0	-0.193	1.189	-0.203	0.848	-0.205	0.675
EQP raw-scale SM=0.3	-0.347	1.235	-0.341	0.858	-0.326	0.690
EQP raw-scale SM=1.0	-0.340	1.224	-0.316	0.858	-0.339	0.690
EQP raw-raw-scale SM=0.0	-0.200	1.186	-0.183	0.834	-0.195	0.677
EQP raw-raw-scale SM=0.3	-0.233	1.226	-0.224	0.863	-0.232	0.685
EQP raw-raw-scale SM=1.0	-0.208	1.228	-0.211	0.858	-0.201	0.699

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

5.18. PROMIS Sleep Disturbance and PROMIS Sleep-related Impairment

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Sleep Disturbance, namely the PROMIS Sleep Disturbance item bank (27 items) and PROMIS Sleep-related Impairment (16 items). Both instruments were scaled such that higher scores represent higher levels of Sleep Disturbance. We excluded 1 participant because of missing responses, leaving a final sample of N=1013. 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 1135 for PROMIS Sleep Disturbance and 80 for PROMIS Sleep-related Impairment. Figure 5.18.1 and Figure 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 Sleep Disturbance and PROMIS Sleep-related Impairment was 0.78. The disattenuated (corrected for unreliabilities) correlation between PROMIS Sleep Disturbance and PROMIS Sleep-related Impairment was 0.81. The correlations between the combined score and the measures were 0.97 and 0.91 for PROMIS Sleep Disturbance and PROMIS Sleep-related Impairment, respectively.

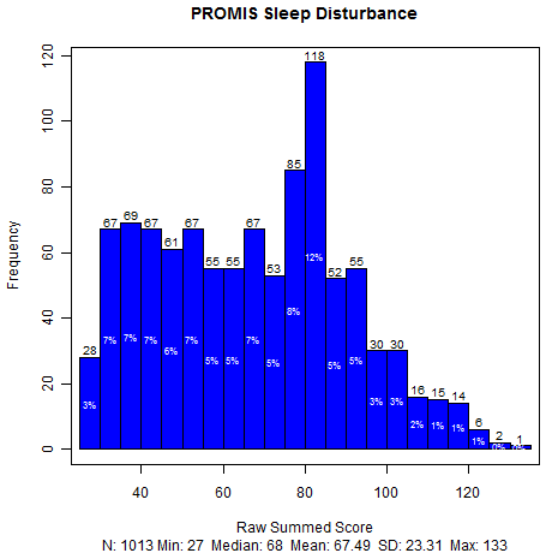


Figure 5.18.1: Raw Summed Score Distribution - PROMIS Sleep Disturbance

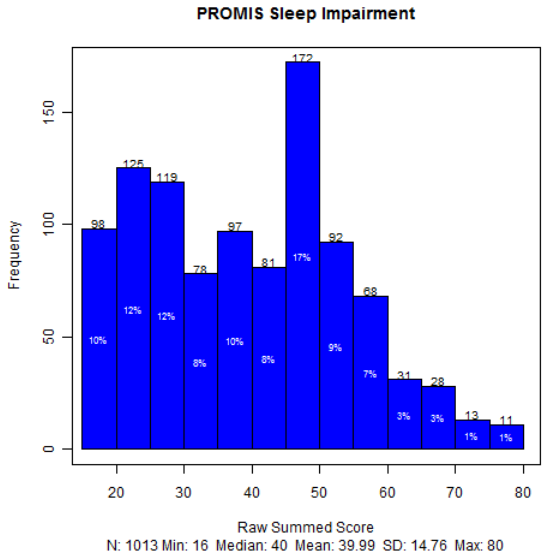


Figure 5.18.2: Raw Summed Score Distribution – PROMIS Sleep-related Impairment

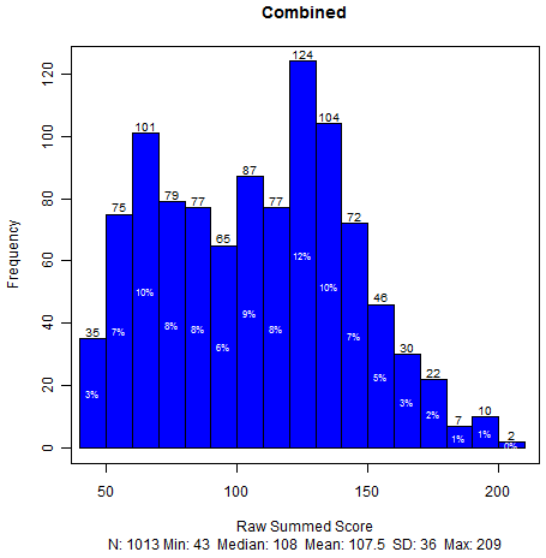


Figure 5.18.3: Raw Summed Score Distribution – Combined

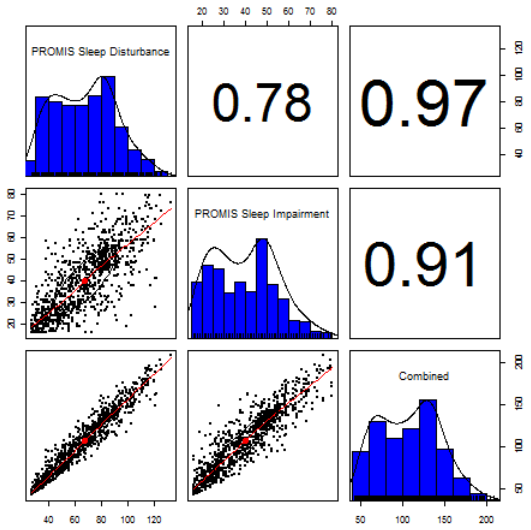


Figure 5.18.4: Scatter Plot Matrix of Raw Summed Scores

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 Sleep Disturbance, Cronbach's alpha internal consistency reliability estimate was 0.964 and adjusted (corrected for overlap) item-total correlations ranged from 0.512 to 0.827. For PROMIS Sleep-related Impairment, alpha was 0.952 and adjusted item-total correlations ranged from 0.56 to 0.853. For the 43 items, alpha was 0.975 and adjusted item-total correlations ranged from 0.51 to 0.808.

Table 5.18.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Sleep Disturbance	27	0.964	0.512	0.694	0.827
PROMIS Sleep-related Impairment	16	0.952	0.560	0.727	0.853
Combined	43	0.975	0.510	0.679	0.808

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 Sleep Disturbance, the fit statistics were as follows: CFI= 0.872, TLI = 0.861, and RMSEA = 0.158. For PROMIS Sleep-related Impairment, CFI = 0.911, TLI = 0.897, and RMSEA

= 0.21. For the 43 items, CFI = 0.848,TLI = 0.84, and RMSEA = 0.133. 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 Sleep Disturbance	27	1013	0.872	0.861	0.158
PROMIS Sleep-related Impairment	16	1013	0.911	0.897	0.210
Combined	43	1013	0.848	0.840	0.133

5.18.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 43 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 43 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 27 PROMIS Sleep Disturbance items served as anchor items to transform the item parameter estimates for the PROMIS Sleep-related Impairment items onto the PROMIS Sleep Disturbance 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 Sleep Disturbance items were constrained to their final bank values, while the PROMIS Sleep-related Impairment items were calibrated, under the constraints imposed by the anchor items.

Table 5.18.3: IRT Linking Constants

	A	B
Mean/Mean	0.978	0.420
Mean/Sigma	1.023	0.401
Haebara	0.962	0.405
Stocking-Lord	1.014	0.381

The item parameter estimates for the PROMIS Sleep-related Impairment items were linked to the PROMIS Sleep Disturbance metric using the transformation constants shown in Table 5.18.3. The PROMIS Sleep-related Impairment item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Sleep Disturbance metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for PROMIS Sleep-related Impairment 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.

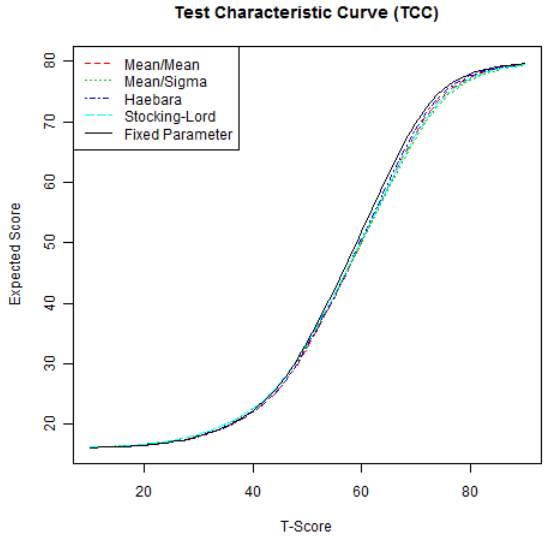


Figure 5.18.5: Test Characteristic Curves (TCC) from Different Linking Methods

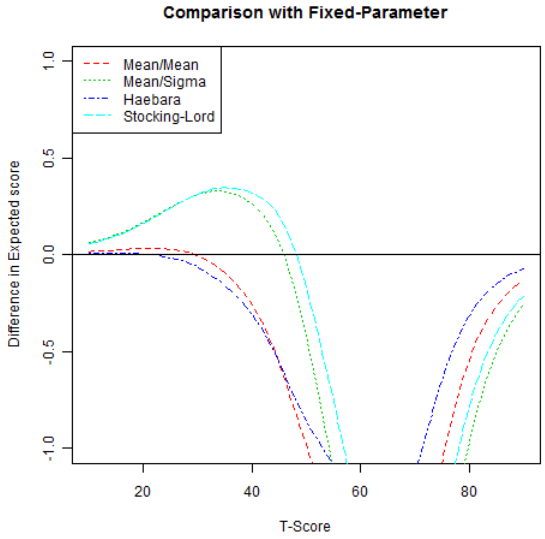


Figure 5.18.6: Difference in Test Characteristic Curves (TCC)

Table 5.18.4 shows the fixed-parameter calibration item parameter estimates for PROMIS Sleep-related Impairment. The marginal reliability estimate for PROMIS Sleep-related Impairment based on the item parameter estimates was 0.941. The marginal reliability estimates for PROMIS Sleep Disturbance and the combined set were 0.964 and 0.977, respectively. The slope parameter estimates for PROMIS Sleep-related Impairment ranged from 1.48 to 3.78 with a mean of 2.44. The slope parameter estimates for PROMIS Sleep Disturbance ranged from 1.19 to 3.66 with a mean of 2.15. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.16.7 displays the scale information functions for PROMIS Sleep Disturbance, PROMIS Sleep-related Impairment, and the combined set of 43. 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 5.18.4: Fixed-Parameter Calibration Item Parameter Estimates

a	cb1	cb2	cb3	cb4	NCAT
1.536	-1.621	-0.228	0.913	1.921	5
1.537	-1.548	-0.075	1.076	2.074	5
1.988	-0.368	0.567	1.470	2.148	5
2.844	-0.111	0.642	1.367	2.091	5
2.815	-0.179	0.629	1.296	1.950	5
2.339	-1.132	0.222	0.960	1.751	5
1.647	-0.362	0.470	1.512	2.489	5
3.436	-0.064	0.641	1.263	1.917	5
3.784	-0.084	0.633	1.243	1.873	5
3.332	-0.124	0.543	1.311	1.933	5
3.468	-0.162	0.556	1.172	1.744	5
3.188	0.190	0.720	1.321	1.854	5
1.480	-1.516	-0.234	0.939	1.915	5
1.572	-1.299	-0.240	0.787	1.733	5
1.954	-0.296	0.553	1.298	2.081	5
2.064	-0.794	0.328	1.082	1.863	5

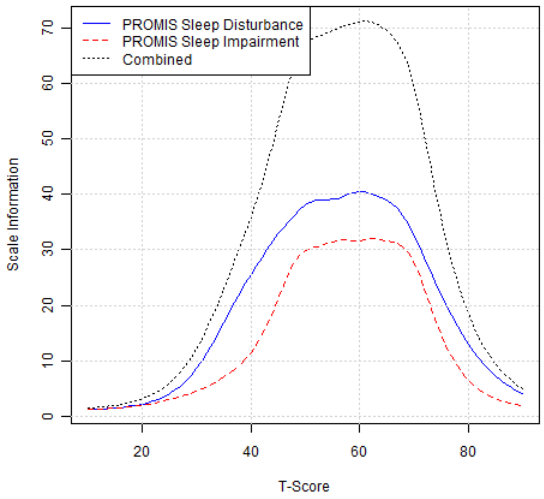


Figure 5.18.7: Comparison of Scale Information Functions

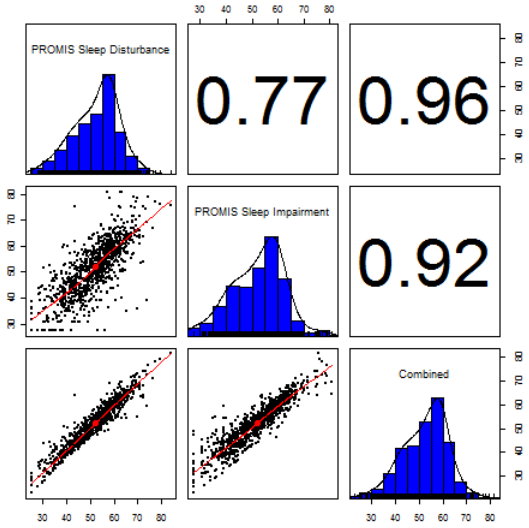


Figure 5.18.8: Comparison of IRT Scaled Scores

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 PROMIS Sleep-related Impairment to a scaled score on PROMIS Sleep Disturbance can be useful. Based on the PROMIS Sleep-related Impairment item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 48 can be used to map simple raw summed scores from PROMIS Sleep-related Impairment to T-score values linked to the PROMIS Sleep Disturbance 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 PROMIS Sleep-related Impairment to a corresponding scaled score on PROMIS Sleep Disturbance by identifying scores on PROMIS Sleep Disturbance that have the same percentile ranks as scores on PROMIS Sleep-related Impairment. 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 PROMIS Sleep-related Impairment to PROMIS Sleep Disturbance. 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 49 and Appendix Table 50 show the equipercentile crosswalk tables. The result shown in Appendix Table 49 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 50 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).

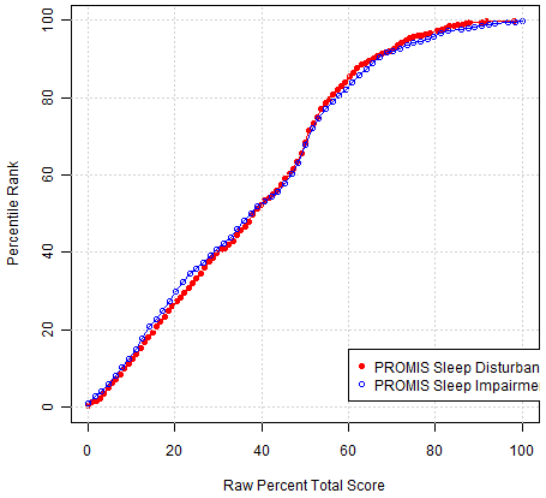


Figure 5.18.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

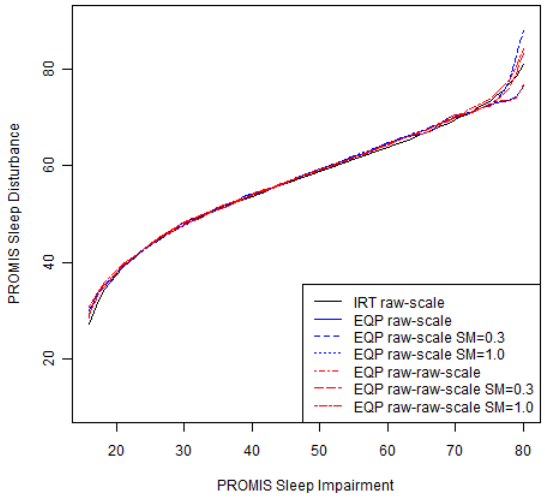


Figure 5.18.10: Equipercentile Linking Functions

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 5.18.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Sleep Disturbance T-scores and PROMIS Sleep-related Impairment 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 (0.766), followed by EQP raw-raw-scale SM=0.0 (0.766). 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 (6.659), followed by EQP raw-scale SM=0.0 (6.668).

Table 5.18.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.766	0.177	6.822	6.821
IRT raw-scale	0.763	0.230	6.751	6.751
EQP raw-scale SM=0.0	0.765	-0.179	6.669	6.668
EQP raw-scale SM=0.3	0.761	-0.274	6.781	6.783
EQP raw-scale SM=1.0	0.761	-0.250	6.785	6.786
EQP raw-raw-scale SM=0.0	0.766	-0.151	6.661	6.659
EQP raw-raw-scale SM=0.3	0.763	-0.232	6.714	6.715
EQP raw-raw-scale SM=1.0	0.763	-0.209	6.698	6.698

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1013) over a large number of replications (i.e., 10,000).

Table 5.18.6 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 Sleep Disturbance 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-scale SM=0.0 produced the smallest standard error, 0.742. That is, the difference between the mean PROMIS Sleep Disturbance T-score and the mean equated PROMIS Sleep-related Impairment T-score based on a similar sample of 75 cases is expected to be around ±1.48 (i.e., 2 × 0.742).

Table 5.18.6: 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	0.166	1.360	0.192	0.933	0.171	0.763
IRT raw-scale	0.243	1.344	0.242	0.936	0.228	0.747
EQP raw-scale SM=0.0	-0.190	1.309	-0.173	0.921	-0.177	0.742
EQP raw-scale SM=0.3	-0.278	1.336	-0.275	0.942	-0.279	0.762
EQP raw-scale SM=1.0	-0.233	1.350	-0.256	0.944	-0.249	0.761
EQP raw-raw-scale SM=0.0	-0.140	1.305	-0.156	0.910	-0.156	0.744
EQP raw-raw-scale SM=0.3	-0.238	1.314	-0.240	0.924	-0.221	0.751
EQP raw-raw-scale SM=1.0	-0.218	1.325	-0.207	0.914	-0.213	0.757

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

5.19. PROMIS Sleep Disturbance and PSQI

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Sleep Disturbance, namely the PROMIS Sleep Disturbance item bank (27 items) and PSQI (7 items). Both instruments were scaled such that higher scores represent higher levels of Sleep Disturbance. Our sample consisted of 1880 participants (1873 participants provided complete responses). 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 135 for PROMIS Sleep Disturbance and 28 for PSQI. 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 Anxiety and Neuro-QoL Anxiety was 0.99. The correlation between PROMIS Sleep Disturbance and PSQI was 0.83. The disattenuated (corrected for unreliabilities) correlation between PROMIS Sleep Disturbance and PSQI was 0.98. The correlations between the combined score and the measures were 1 and 0.88 for PROMIS Sleep Disturbance and PSQI, respectively.

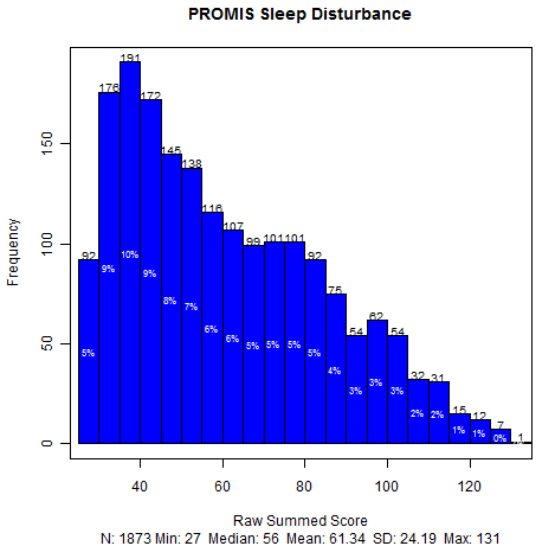


Figure 5.19.1: Raw Summed Score Distribution - PROMIS Sleep Disturbance

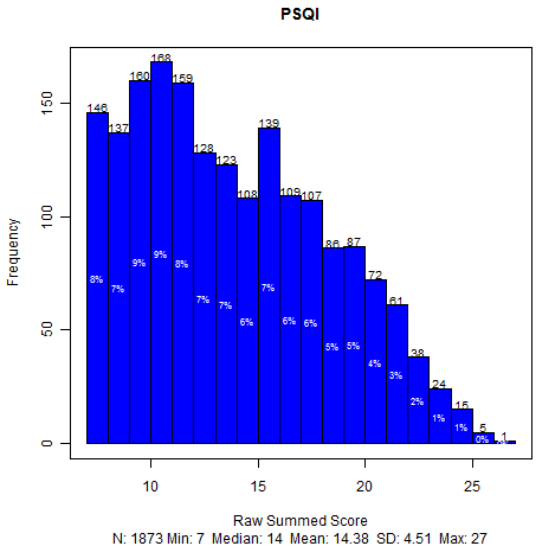


Figure 5.19.2: Raw Summed Score Distribution - PSQI

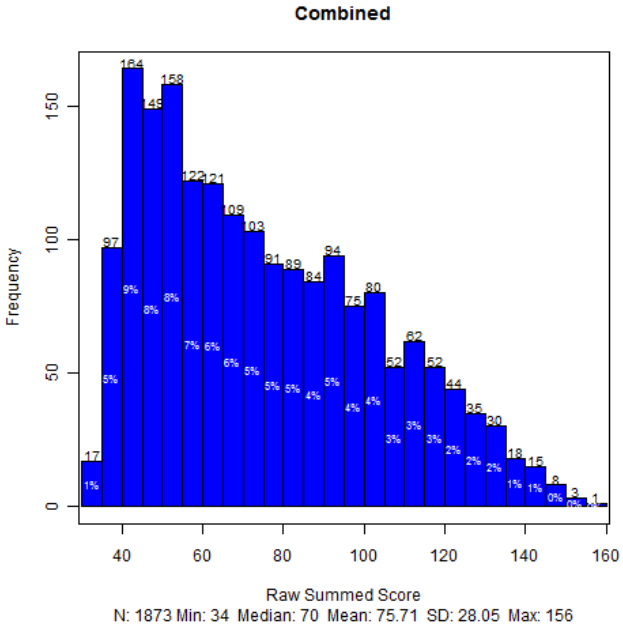


Figure 5.19.3: Raw Summed Score Distribution – Combined

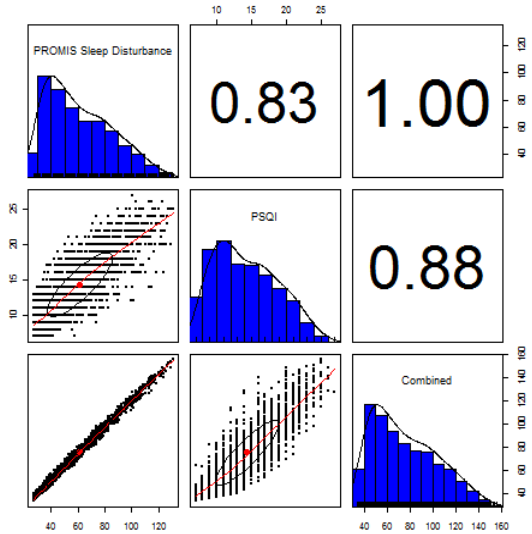


Figure 5.19.4: Scatter Plot Matrix of Raw Summed Scores

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 Sleep Disturbance, Cronbach's alpha internal consistency reliability estimate was 0.969 and adjusted (corrected for overlap) item-total correlations ranged from 0.553 to 0.869. For PSQI, alpha was 0.736 and adjusted item-total correlations ranged from 0.274 to 0.686. For the 34 items, alpha was 0.969 and adjusted item-total correlations ranged from 0.25 to 0.869.

Table 5.19.1: Classic Item Analysis

	No.	Alpha	min.r	mean.r	max.r
PROMIS Sleep Disturbance	27	0.969	0.553	0.719	0.869
PSQI	7	0.736	0.274	0.489	0.686
Combined	34	0.969	0.250	0.688	0.869

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 Sleep Disturbance, the fit statistics were as follows: CFI = 0.925, TLI = 0.919, and RMSEA = 0.145. For PSQI, CFI = 0.961, TLI= 0.942, and RMSEA = 0.111. For the 34 items,

CFI = 0.928, TLI = 0.924, and RMSEA = 0.124. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.19.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Sleep	27	1880	0.925	0.919	0.145
PSQI	7	1880	0.961	0.942	0.111
Combined	34	1880	0.928	0.924	0.124

5.19.4. 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 27 PROMIS Sleep Disturbance items served as anchor items to transform the item parameter estimates for the PSQI items onto the PROMIS Sleep Disturbance 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 Sleep Disturbance items were constrained to their final bank values, while the PSQI items were calibrated, under the constraints imposed by the anchor items.

Table 5.19.3: IRT Linking Constants

	A	B
Mean/Mean	1.094	0.076
Mean/Sigma	1.084	0.083
Haebara	1.085	0.084
Stocking-Lord	1.084	0.083

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

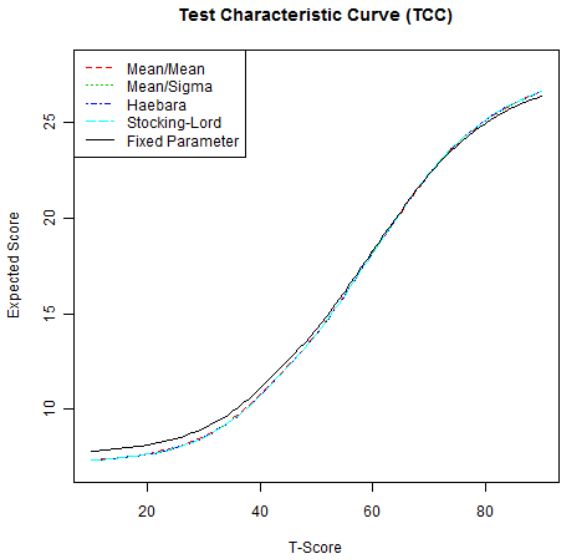


Figure 5.19.5: Test Characteristic Curves (TCC) from Different Linking Methods

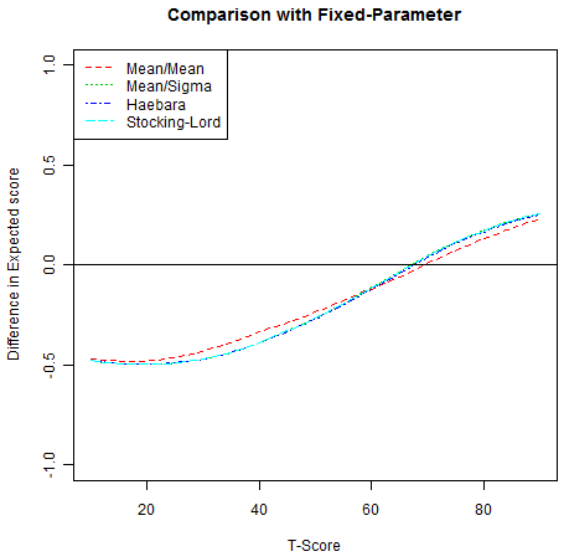


Figure 5.19.6: Difference in Test Charactertistic Curves (TCC)

Table 5.19.4 shows the fixed-parameter calibration item parameter estimates for PSQI. The marginal reliability estimate for PSQI based on the item parameter estimates was 0.825. The marginal reliability estimates for PROMIS Sleep Disturbance and the combined set were 0.964 and 0.969, respectively. The slope parameter estimates for PSQI ranged from 0.276 to 3.23 with a mean of 1.6. The slope parameter estimates for PROMIS Sleep Disturbance ranged from 1.19 to 3.66 with a mean of 2.15. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.19.7 displays the scale information functions for PROMIS Sleep Disturbance, PSQI, and the combined set of 34. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.19.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.19.4: Fixed-Parameter Calibration Item Parameter Estimates for PSQI

a	cb1	cb2	cb3	NCAT
1.157	-0.116	1.108	2.272	4
2.082	-1.740	0.652	2.460	4
2.263	-0.448	0.532	1.310	4
1.276	-0.843	1.574	3.320	4
0.276	-0.546	0.267	0.644	4
3.226	-1.052	0.472	1.707	4
0.890	0.949	1.472	2.005	4

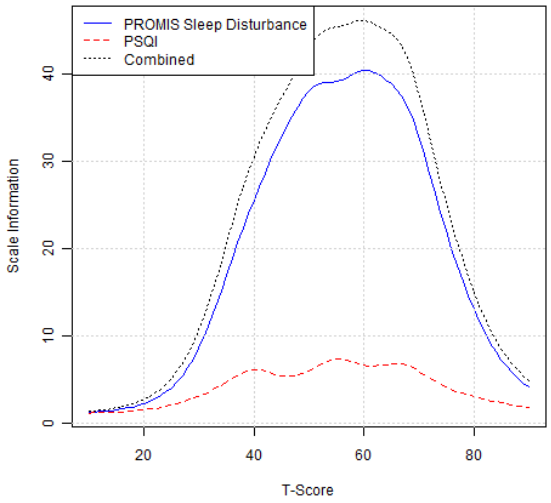


Figure 5.19.7: Comparison of Scale Information Functions

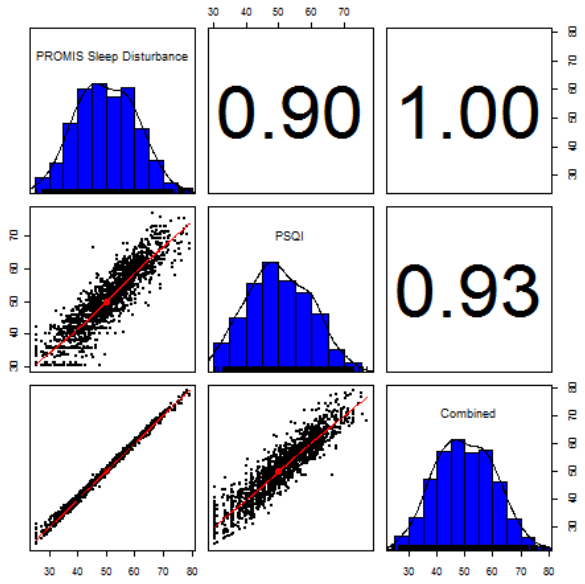


Figure 5.19.8: Comparison of IRT Scaled Scores

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 PSQI to a scaled score on PROMIS Sleep Disturbance can be useful. Based on the PSQI item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 51 can be used to map simple raw summed scores from PSQI to T-score values linked to the PROMIS Sleep Disturbance 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 PSQI to a corresponding scaled score on PROMIS Sleep Disturbance by identifying scores on PROMIS Sleep Disturbance that have the same percentile ranks as scores on PSQI. 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 PSQI to PROMIS Sleep Disturbance. 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 52 and Appendix Table 53 show the equipercentile crosswalk tables. The result shown in Appendix Table 52 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 53 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).

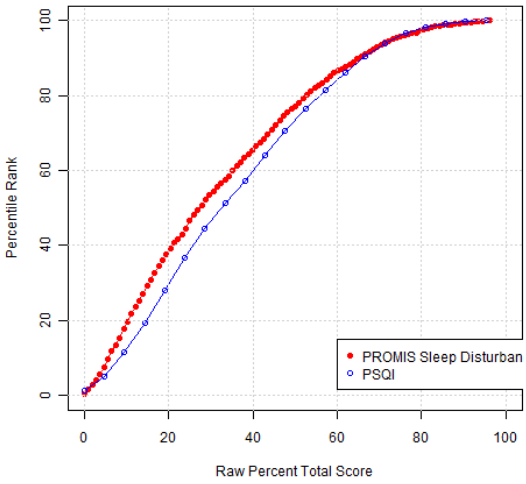


Figure 5.19.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

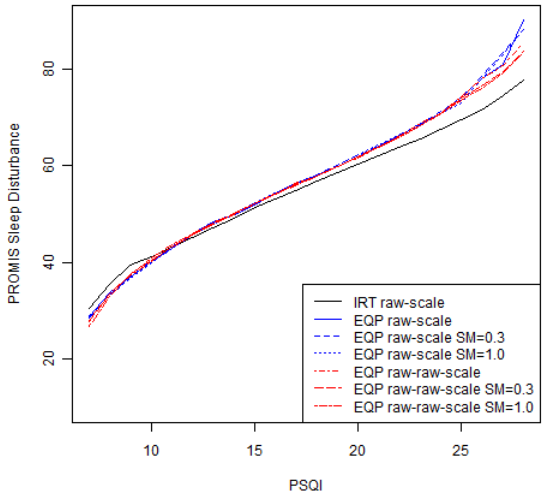


Figure 5.19.10: Equipercentile Linking Functions

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 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 5.19.10, 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. 5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Sleep Disturbance T-scores and PSQI 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, IRT pattern scoring produced the best result (0.905), followed by IRT raw-scale (0.826). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (4.475), followed by IRT raw-scale (5.948).

Table 5.19.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.905	0.021	4.476	4.475
IRT raw-scale	0.826	0.515	5.927	5.948
EQP raw-scale SM=0.0	0.825	0.022	6.168	6.166
EQP raw-scale SM=0.3	0.825	0.038	6.186	6.184
EQP raw-scale SM=1.0	0.825	0.073	6.193	6.192
EQP raw-raw-scale SM=0.0	0.824	0.079	6.171	6.170
EQP raw-raw-scale SM=0.3	0.825	0.059	6.153	6.152
EQP raw-raw-scale SM=1.0	0.823	0.066	6.174	6.173

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1873) over a large number of replications (i.e., 10,000).

Table 5.19.6 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 Sleep Disturbance 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, 0.506. That is, the difference between the mean PROMIS Sleep Disturbance T-score and the mean equated PSQI T-score based on a similar sample of 75 cases is expected to be around
±1.01(i.e., 2 × 0.506).

Table 5.19.6: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	0.045	0.894	0.007	0.624	0.023	0.506
IRT raw-scale	0.518	1.181	0.518	0.833	0.514	0.672
EQP raw-scale SM=0.0	0.012	1.220	0.019	0.864	0.011	0.696
EQP raw-scale SM=0.3	0.046	1.225	0.042	0.863	0.030	0.704
EQP raw-scale SM=1.0	0.077	1.217	0.060	0.880	0.066	0.709

EQP raw-raw-scale SM=0.0	0.064	1.221	0.083	0.863	0.093	0.702
EQP raw-raw-scale SM=0.3	0.050	1.216	0.078	0.854	0.061	0.687
EQP raw-raw-scale SM=1.0	0.073	1.221	0.068	0.856	0.061	0.686

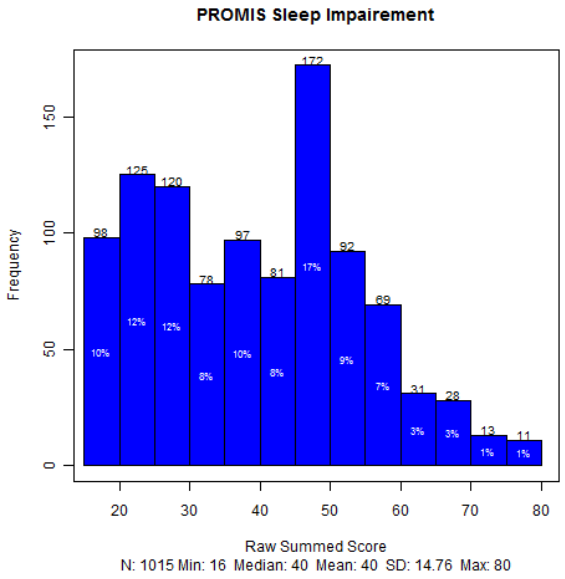
Examining a number of linking studies in the current project revealed that the two linking methods (IRT and equipercentile) in general produced highly comparable results. Some noticeable discrepancies were observed (albeit rarely) in some extreme score levels where data were sparse. Model-based approaches 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 51) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.20. PROMIS Sleep-related Impairment and Neuro-QoL Sleep Disturbance

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Sleep-related Impairment, namely the PROMIS Sleep-related Impairment item bank (16 items) and Neuro-QoL Sleep Disturbance (8 items). Both instruments were scaled such that higher scores represent higher levels of Sleep-related Impairment. We excluded 1 participant because of missing responses, leaving a final sample of N=1015. 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 80 for PROMIS Sleep-related Impairment and 40 for Neuro-QoL Sleep Disturbance. 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 Sleep Impairment and Neuro-QoL Sleep Disturbance was 0.81. The disattenuated (corrected for unreliabilities) correlation between PROMIS Sleep-related Impairment and Neuro-QoL Sleep Disturbance was 0.89. The correlations between the combined score and the measures were 0.98 and 0.91 for PROMIS Sleep-related Impairment and Neuro-QoL Sleep Disturbance, respectively.



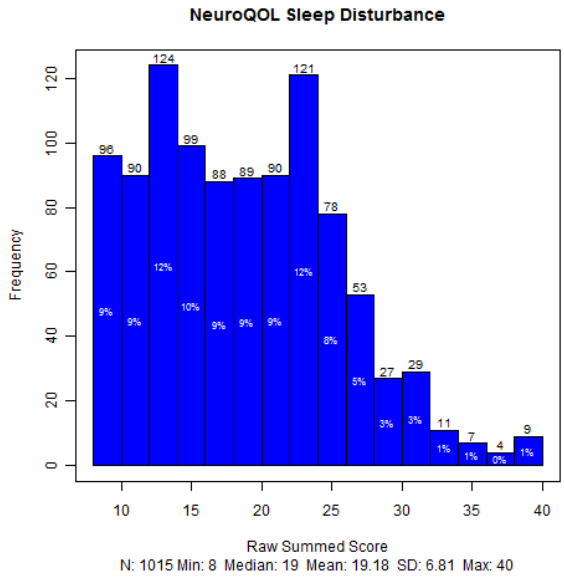


Figure 5.20.1: Raw Summed Score Distribution - PROMIS Sleep-related Impairment

Figure 5.20.2: Raw Summed Score Distribution – Neuro-QOL Sleep Disturbance

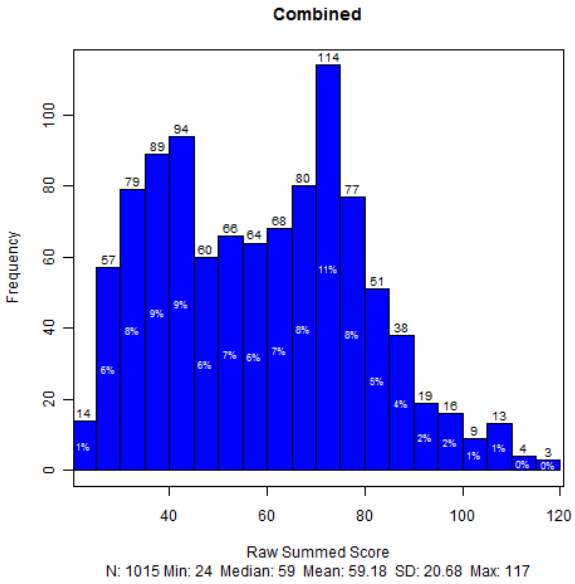


Figure 5.20.3: Raw Summed Score Distribution – Combined

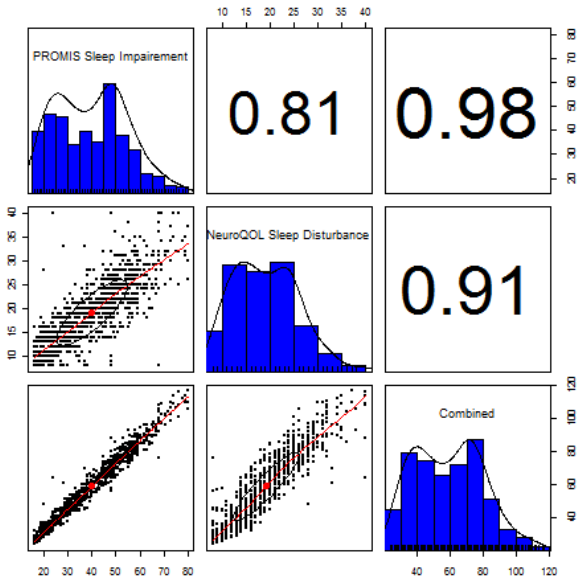


Figure 5.20.4: Scatter Plot Matrix of Raw Summed Scores

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 Sleep Impairment, Cronbach's alpha internal consistency reliability estimate was 0.952 and adjusted (corrected for overlap) item-total correlations ranged from 0.557 to 0.852. For Neuro-QoL Sleep Disturbance, alpha was 0.878 and adjusted item- total correlations ranged from 0.507 to 0.728. For the 24 items, alpha was 0.96 and adjusted item-total correlations ranged from 0.542 to 0.849.

Table 5.20.1: Classical Item Analysis

	No.	Alpha	min.r	mean.r	max.r
PROMIS Sleep-related Impairment	16	0.952	0.557	0.727	0.852
Neuro-QoL Sleep Disturbance	8	0.878	0.507	0.640	0.728
Combined	24	0.960	0.542	0.694	0.849

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 Sleep-related Impairment, the fit statistics were as follows: CFI = 0.909, TLI = 0.895, and RMSEA = 0.212. For Neuro-QoL Sleep Disturbance, CFI = 0.949, TLI = 0.929, and RMSEA = 0.148. For the 24 items, CFI = 0.899, TLI = 0.889, and RMSEA = 0.156. 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 Sleep-related Impairment	16	1015	0.909	0.895	0.212
Neuro-QoL Sleep Disturbance	8	1015	0.949	0.929	0.148
Combined	24	1015	0.899	0.889	0.156

5.20.4. 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 16 PROMIS Sleep-related Impairment items served as anchor items to transform the item parameter estimates for the Neuro-QoL Sleep Disturbance items onto the PROMIS Sleep-related Impairment 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 Sleep-related Impairment items were con- strained to their final bank values, while the Neuro-QoL Sleep Disturbance items were calibrated, under the constraints imposed by the anchor items.

Table 5.20.3: IRT Linking Constants

	A	B
Mean/Mean	1.148	0.439
Mean/Sigma	1.186	0.424
Haebara	1.043	0.529
Stocking-Lord	1.178	0.411

The item parameter estimates for the Neuro-QoL Sleep Disturbance items were linked to the PROMIS Sleep-related Impairment metric using the transformation constants shown in Table 5.20.3. The Neuro-QoL Sleep Disturbance item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Sleep-related Impairment metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for

Neuro-QoL Sleep Disturbance 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.

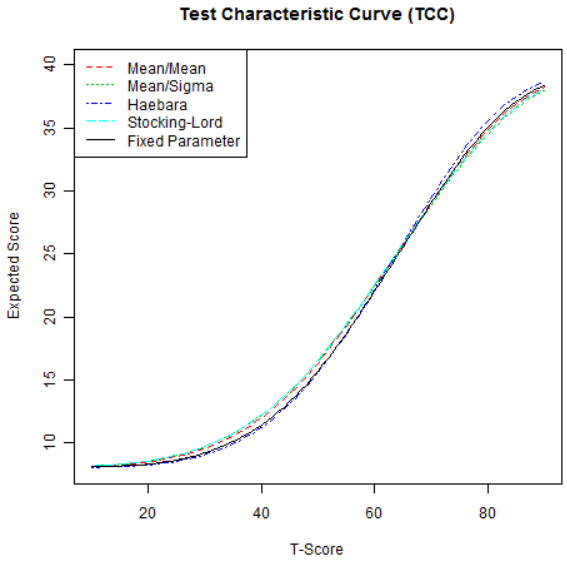


Figure 5.20.5: Test Characteristic Curves (TCC) from Different Linking Methods

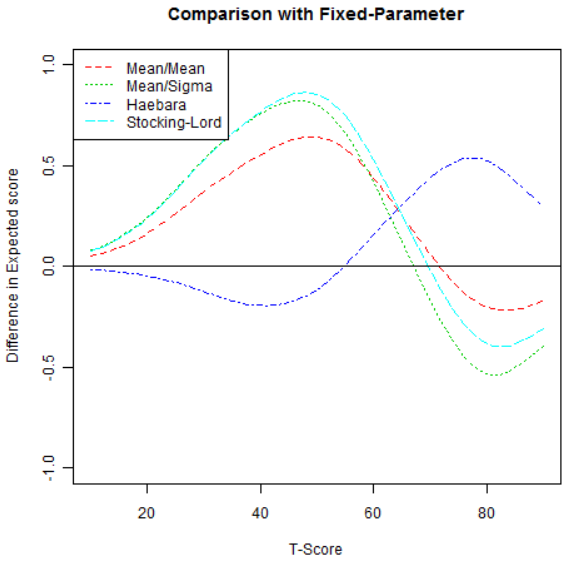


Figure 5.20.6: Difference in Test Characteristic Curves (TCC)

Table 5.20.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QoL Sleep Disturbance. The marginal reliability estimate for Neuro-QoL Sleep Disturbance based on the item parameter estimates was 0.824. The marginal reliability estimates for PROMIS Sleep-related Impairment and the combined set were 0.945 and 0.956, respectively. The slope parameter estimates for Neuro-QOL Sleep Disturbance ranged from 1.31 to 2.29 with a mean of 1.63. The slope parameter estimates for PROMIS Sleep-related Impairment ranged from 1.18 to 4.82 with a mean of 2.59. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.20.7 displays the scale information functions for PROMIS Sleep-related Impairment, Neuro-QoL Sleep Disturbance, and the combined set of 24. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.20.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.20.4: Fixed-Parameter Calibration Item Parameter Estimates for Neuro-QoL Sleep Disturbance

a	cb1	cb2	cb3	cb4	NCAT
1.358	-0.996	0.323	1.612	2.795	5
1.370	-0.998	0.202	1.634	2.808	5
1.760	-1.631	-0.204	1.206	2.405	5
1.395	0.163	1.316	2.303	3.397	5
1.557	-0.560	0.554	1.672	2.771	5
1.309	0.173	1.052	2.231	3.315	5
2.294	0.479	1.138	1.943	2.720	5
1.968	0.131	0.899	1.919	2.756	5

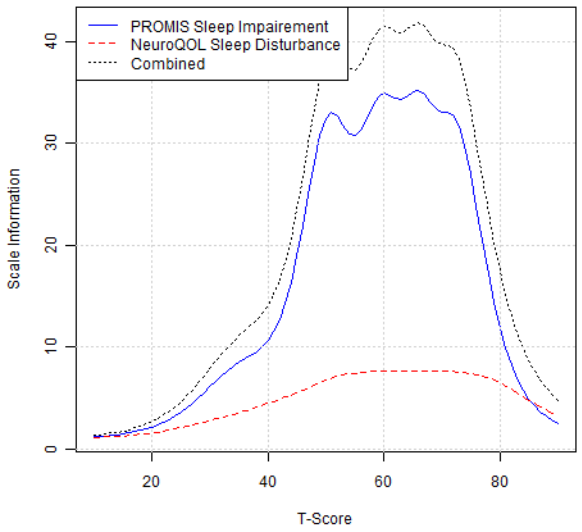


Figure 5.20.7: Comparison of Scale Information Functions

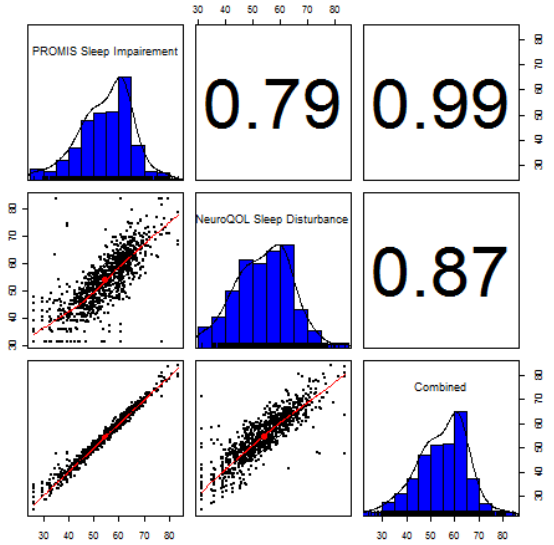


Figure 5.20.8: Comparison of IRT Scaled Scores

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 Sleep Disturbance to a scaled score on PROMIS Sleep-related Impairment can be useful. Based on the Neuro-QoL Sleep Disturbance item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 54 can be used to map simple raw summed scores from Neuro-QoL Sleep Disturbance to T-score values linked to the PROMIS Sleep-related Impairment 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 Sleep Disturbance to a corresponding scaled score on PROMIS Sleep-related Impairment by identifying scores on PROMIS Sleep-related Impairment that have the same percentile ranks as scores on Neuro-QoL Sleep Disturbance. 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 Sleep Disturbance to PROMIS Sleep-related Impairment. 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 55 and 56 show the equipercentile crosswalk tables. The result shown in Appendix Table 55 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 56 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).

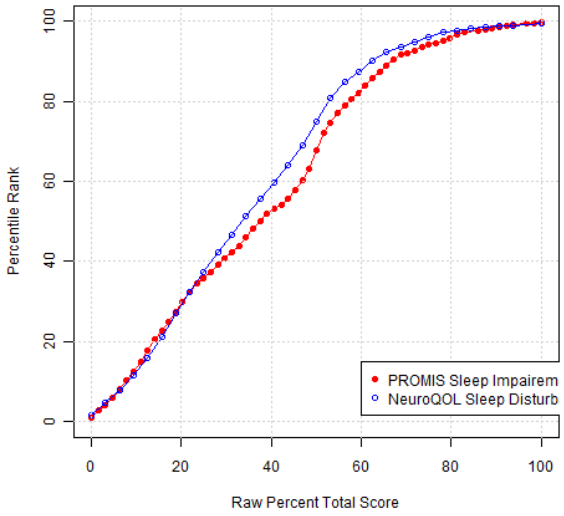


Figure 5.20.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

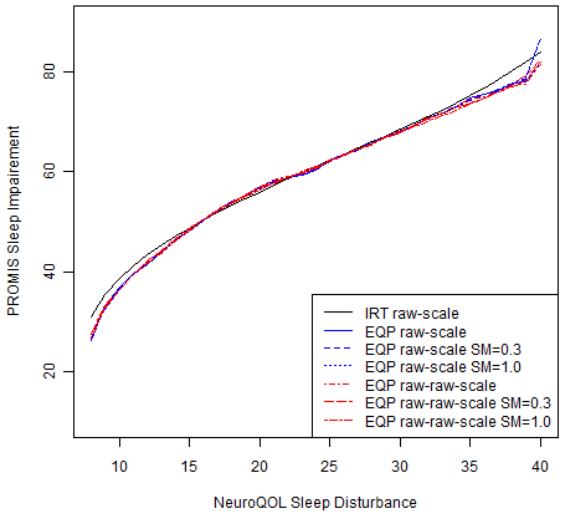


Figure 5.20.9: Equipercentile Linking Functions

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 5.20.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Sleep-related Impairment T-scores and Neuro-QoL Sleep Disturbance scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.20.10), 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 (0.791), followed by EQP raw-raw-scale SM=0.0 (0.79). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (6.837), followed by IRT raw-scale (6.905).

Table 5.20.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.791	0.716	6.803	6.837
IRT raw-scale	0.787	0.908	6.848	6.905
EQP raw-scale SM=0.0	0.787	1.469	7.072	7.219
EQP raw-scale SM=0.3	0.785	1.425	7.136	7.274
EQP raw-scale SM=1.0	0.786	1.434	7.113	7.253
EQP raw-raw-scale SM=0.0	0.790	1.410	7.001	7.138
EQP raw-raw-scale SM=0.3	0.789	1.437	7.005	7.147
EQP raw-raw-scale SM=1.0	0.788	1.449	7.031	7.176

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1015) over a large number of replications (i.e., 10,000).

Table 5.20.6 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 Sleep-related Impairment 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, 0.746. That is, the difference between the mean PROMIS Sleep-related Impairment T-score and the mean equated Neuro-QoL Sleep Disturbance T-score based on a similar sample of 75 cases is expected to be around ±1.49 (i.e., 2 × 0.746).

Table 5.20.6: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	0.722	1.335	0.715	0.943	0.720	0.746
IRT raw-scale	0.914	1.374	0.902	0.945	0.915	0.760
EQP raw-scale SM=0.0	1.457	1.419	1.480	0.984	1.465	0.795
EQP raw-scale SM=0.3	1.414	1.404	1.403	0.988	1.425	0.793
EQP raw-scale SM=1.0	1.429	1.401	1.423	0.982	1.432	0.789
EQP raw-raw-scale SM=0.0	1.415	1.396	1.424	0.979	1.409	0.771
EQP raw-raw-scale SM=0.3	1.423	1.362	1.430	0.959	1.446	0.785
EQP raw-raw-scale SM=1.0	1.462	1.391	1.442	0.980	1.432	0.784

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

5.21. PROMIS Sleep-related Impairment and PSQI

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Sleep Impairment, namely the PROMIS Sleep-related Impairment item bank (16 items) and PSQI (7 items). Both instruments were scaled such that higher scores represent higher levels of Sleep Impairment. We excluded 1 participant because of missing responses, leaving a final sample of N=1878. 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.

5.21.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 80 for PROMIS Sleep-related Impairment and 28 for PSQI. Figures 5.21.1 and 5.21.2 graphically display the raw summed score distributions of the two measures. Figure 5.21.3 shows the distribution for the combined. Figure 5.21.4 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 Sleep- related Impairment and PSQI was 0.72. The disattenuated (corrected for unreliabilities) correlation between PROMIS Sleep-related Impairment and PSQI was 0.86. The correlations between the combined score and the measures were 0.98 and 0.84 for PROMIS Sleep- related Impairment and PSQI, respectively.

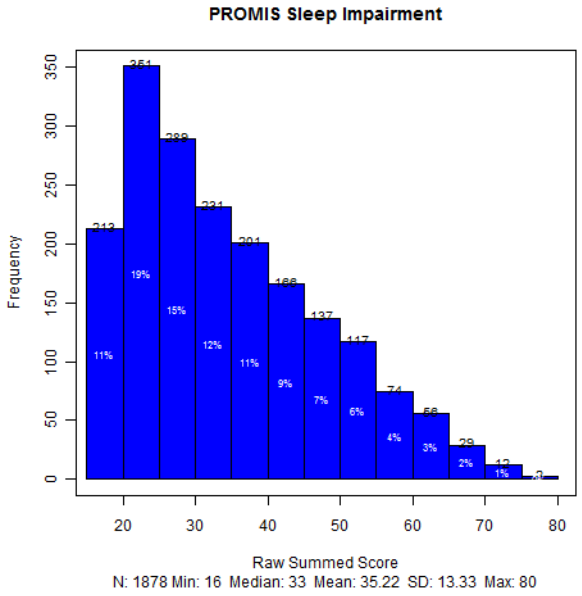


Figure 5.21.1: Raw Summed Score Distribution - PROMIS Sleep-related Impairment

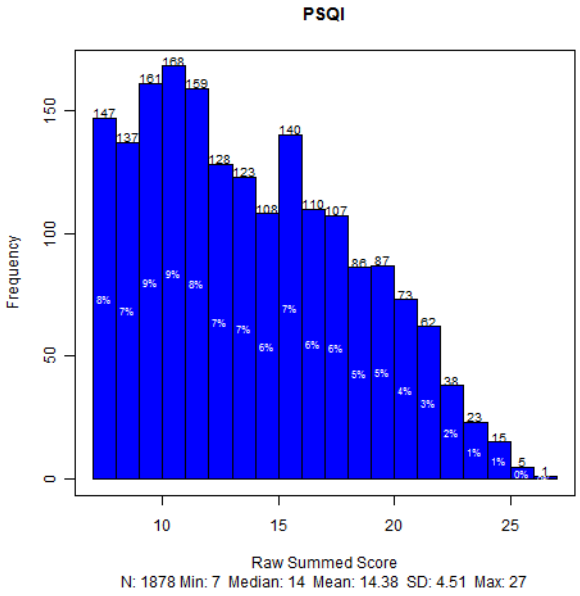


Figure 5.21.2: Raw Summed Score Distribution - PSQI

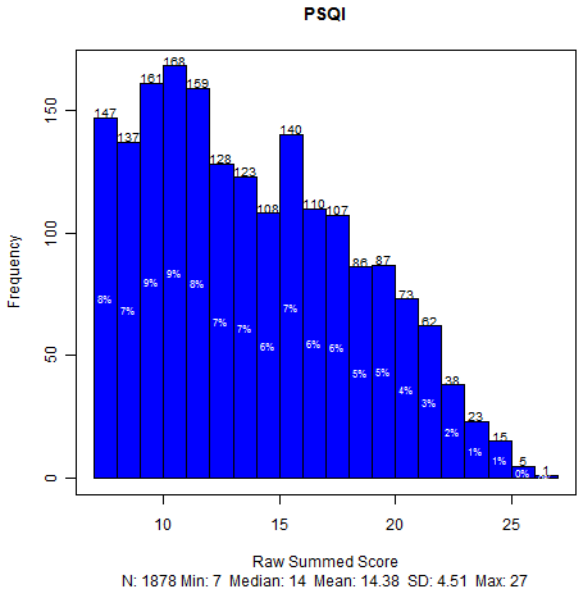


Figure 5.21.3: Raw Summed Score Distribution – Combined

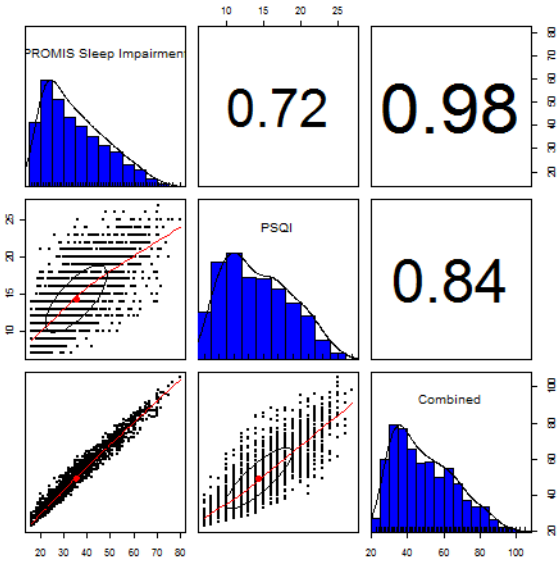


Figure 5.21.4: Scatter Plot Matrix of Raw Summed Scores

5.21.2. 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 Sleep- related Impairment, Cronbach's alpha internal consistency reliability estimate was 0.95 and adjusted (corrected for overlap) item-total correlations ranged from 0.523 to 0.828. For PSQI, alpha was 0.736 and adjusted item-total correlations ranged from 0.274 to 0.686. For the 23 items, alpha was 0.947 and adjusted item-total correlations ranged from 0.288 to 0.822.

Table 5.21.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
PROMIS Sleep- related Impairment	16	0.950	0.523	0.726	0.828
PSQI	7	0.736	0.274	0.489	0.686
Combined	23	0.947	0.288	0.659	0.822

5.21.3. 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 5.21.2 summarizes the model fit statistics. For PROMIS Sleep- related Impairment, the fit statistics were as follows: CFI = 0.941, TLI = 0.932, and RMSEA = 0.185. For PSQI, CFI = 0.961, TLI = 0.942, and RMSEA = 0.111. For the 23 items, CFI = 0.931, TLI = 0.924, and RMSEA = 0.144. The main interest of the current analysis is whether the combined measure is essential unidimensional.

Table 5.21.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Sleep- related Impairment	16	1880	0.941	0.932	0.185
PSQI	7	1880	0.961	0.942	0.111
Combined	23	1880	0.931	0.924	0.144

5.21.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 23 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 23 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 16 PROMIS Sleep- related Impairment items served as anchor items to transform the item parameter estimates for the PSQI items onto the PROMIS Sleep- related Impairment 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 5.21.3 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 Sleep- related Impairment items were constrained to their final bank values, while the PSQI items were calibrated, under the constraints imposed by the anchor items.

Table 5.21.3: IRT Linking Constants

	A	B
Mean/Mean	1.020	0.076
Mean/Sigma	1.019	0.076
Haebara	1.020	0.076
Stocking-Lord	1.021	0.079

The item parameter estimates for the PSQI items were linked to the PROMIS Sleep- related Impairment metric using the transformation constants shown in Table 5.21.3. The PSQI item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Sleep- related Impairment metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for PSQI as shown in Figure 5.21.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.21.6 displays the differences on the vertical axis.

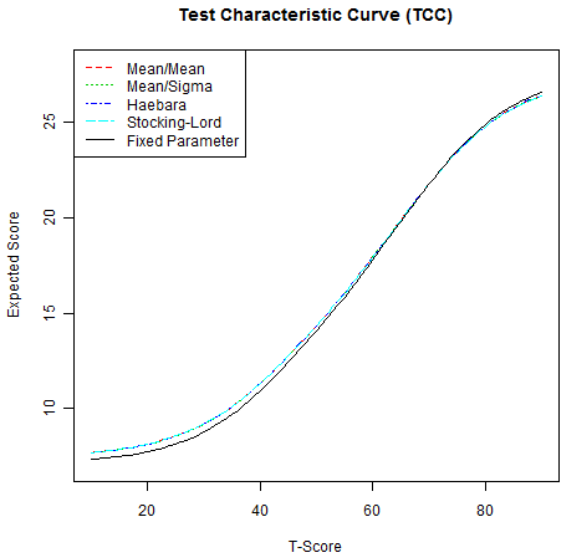


Figure 5.21.5: Test Characteristic Curves (TCC) from Different Linking Methods

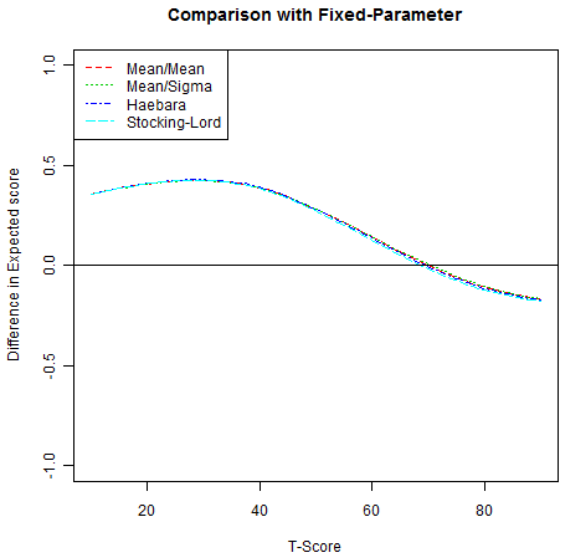


Figure 5.21.6: Difference in Test Characteristic Curves (TCC)

Table 5.21.4 shows the fixed-parameter calibration item parameter estimates for PSQI. The marginal reliability estimate for PSQI based on the item parameter estimates was 0.773. The marginal reliability estimates for PROMIS Sleep- related Impairment and the combined set were 0.945 and 0.955, respectively. The slope parameter estimates for PSQI ranged from 0.584 to 2.2 with a mean of 1.35. The slope parameter estimates for PROMIS Sleep- related Impairment ranged from 1.18 to 4.82 with a mean of 2.59. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.21.7 displays the scale information functions for PROMIS Sleep- related Impairment, PSQI, and the combined set of 23. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.21.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.19.4: Fixed-Parameter Calibration Item Parameter Estimates for PSQI

a	cb1	cb2	cb3	NCAT
0.892	-0.110	1.319	2.693	4
1.614	-1.883	0.710	2.729	4
1.286	-0.522	0.676	1.643	4
2.202	-0.632	1.209	2.451	4
0.584	0.518	0.733	0.869	4
2.100	-1.118	0.532	1.895	4
0.741	1.073	1.666	2.285	4

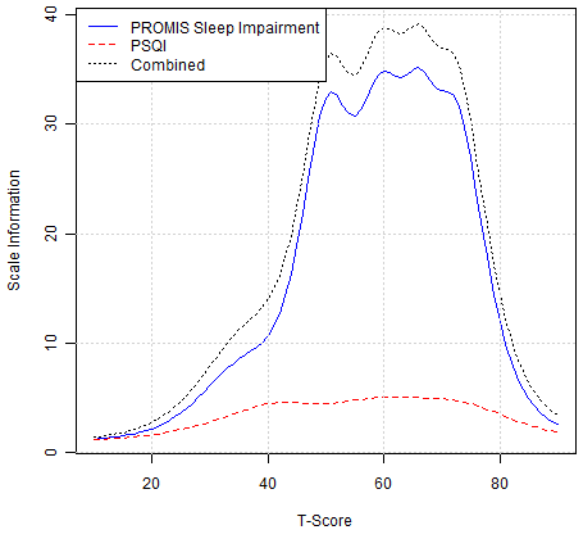


Figure 5.21.7: Comparison of Scale Information Functions

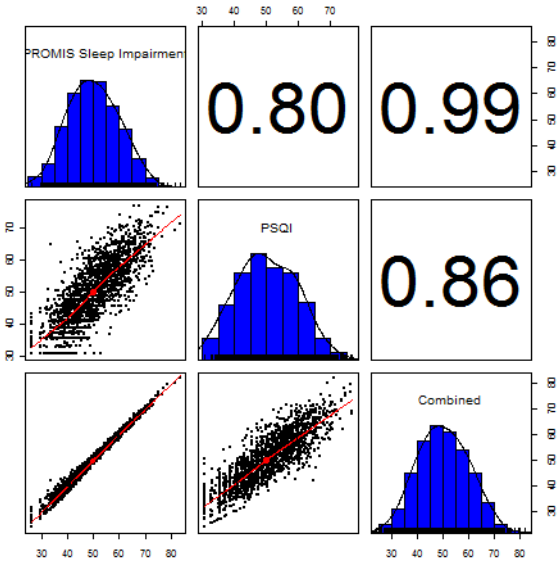


Figure 5.21.8: Comparison of IRT Scaled Scores

5.21.5. 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 PSQI to a scaled score on PROMIS Sleep-related Impairment can be useful. Based on the PSQI item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 57 can be used to map simple raw summed scores from PSQI to T-score values linked to the PROMIS Sleep- related Impairment. 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.

5.21.6. Equipercentile Linking

We mapped each raw summed score point on PSQI to a corresponding scaled score on PROMIS Sleep Disturbance by identifying scores on PROMIS Sleep Disturbance that have the same percentile ranks as scores on PSQI. 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 5.21.9 displays the cumulative distribution functions of the measures. Figure 5.21.10 shows the equipercentile linking functions based on raw summed

scores from PSQI to PROMIS Sleep- related Impairment. 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 58 and 59 show the equipercentile crosswalk tables. The result shown in Appendix Table 58 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 59 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).

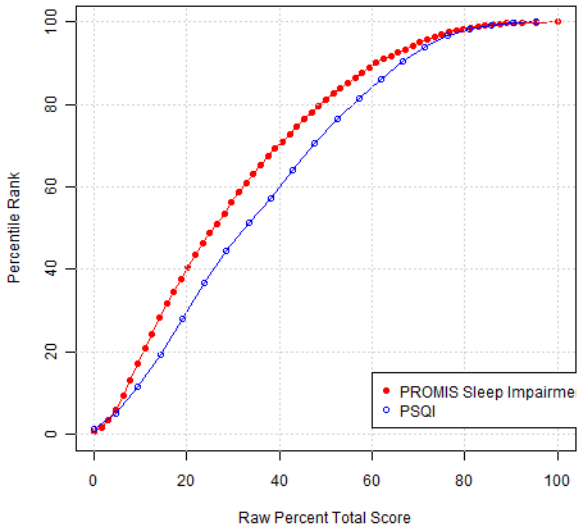


Figure 5.21.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

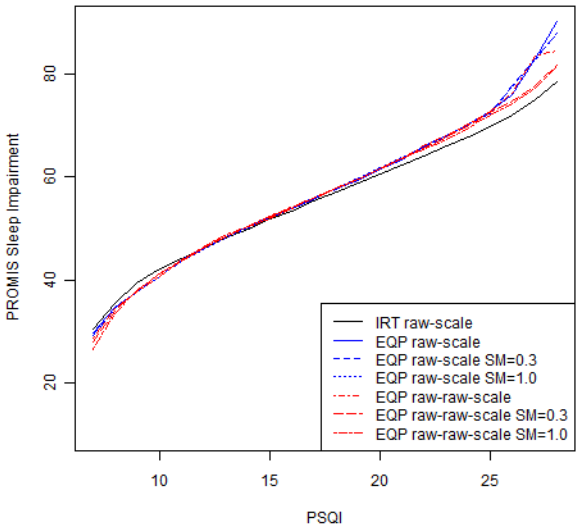


Figure 5.21.10: Equipercentile Linking Functions

5.21.7. 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 5.21.10, 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 5.21.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Sleep- related Impairment T-scores and PSQI scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.21.10), 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 (0.797), followed by IRT raw-scale (0.718). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (6.205), followed by IRT raw-scale (7.135).

Table 5.21.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.797	0.058	6.206	6.205
IRT raw-scale	0.718	0.183	7.135	7.135
EQP raw-scale SM=0.0	0.717	0.059	7.444	7.442
EQP raw-scale SM=0.3	0.717	0.076	7.467	7.465
EQP raw-scale SM=1.0	0.717	0.065	7.479	7.477
EQP raw-raw-scale	0.717	0.094	7.469	7.468
EQP raw-raw-scale	0.717	0.066	7.453	7.451
EQP raw-raw-scale	0.715	0.089	7.496	7.495

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1878) over a large number of replications (i.e., 10,000).

Table 5.21.6 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 Sleep-related Impairment 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, 0.701. That is, the difference between the mean PROMIS Sleep-related Impairment T-score and the mean equated PSQI T-score based on a similar sample of 75 cases is expected to be around ±1.4 (i.e., 2 × 0.701).

Table 5.21.6: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	0.042	1.231	0.063	0.865	0.058	0.701
IRT raw-scale	0.202	1.423	0.173	1.008	0.188	0.809
EQP raw-scale SM=0.0	0.047	1.486	0.047	1.032	0.064	0.840
EQP raw-scale SM=0.3	0.088	1.472	0.096	1.040	0.067	0.842
EQP raw-scale SM=1.0	0.072	1.492	0.052	1.046	0.063	0.848
EQP raw-raw-scale SM=0.0	0.098	1.493	0.100	1.049	0.100	0.848
EQP raw-raw-scale SM=0.3	0.099	1.465	0.073	1.052	0.079	0.838
EQP raw-raw-scale SM=1.0	0.085	1.496	0.084	1.043	0.098	0.835

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

5.22. PROMIS Satisfaction with Social Roles and Activities (v2.0) and PROMIS Satisfaction with Participation in Discretionary Social Activities (v1.0)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Social Functioning, namely the PROMIS Satisfaction with Social Roles and Activities v2.0 (P Social RAV2) item bank (44 items) and the PROMIS Satisfaction with Participation in Discretionary Social Activities v1.0 (P Partic DSAV1) item bank (12 items). Both instruments were scaled such that higher scores represent higher levels of Social. We did not exclude any participants because of missing responses, leaving a final sample of N=1007. 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.

5.22.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 220 for P Social RAV2 and 60 for P Partic DSAV1. Figures 5.22.1 and 5.22.2 graphically display the raw summed score distributions of the two measures. Figure 5.22.3 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 P Social RAV2 and P Partic DSAV1 was 0.86. The disattenuated (corrected for unreliabilities) correlation between P Social RAV2 and P Partic DSAV1 was 0.89. The correlations between the combined score and the measures were 0.99 and 0.91 for P Social RAV2 and P Partic DSAV1, respectively.

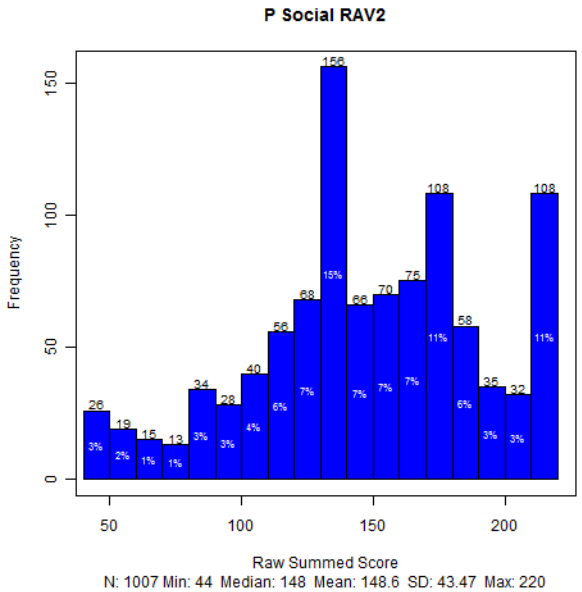


Figure 5.22.1: Raw Summed Score Distribution – PROMIS Satisfaction w/ Social Roles and Activities

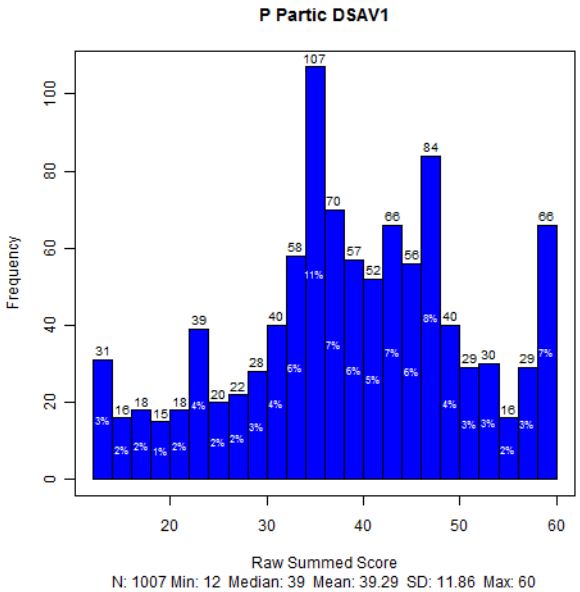


Figure 5.22.2: Raw Summed Score Distribution – PROMIS Satisfaction w/ Participation in Discretionary Social Activities

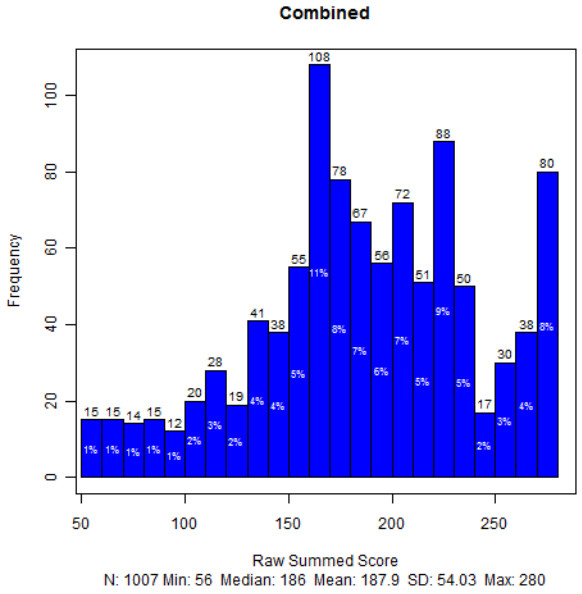


Figure 5.22.3: Raw Summed Score Distribution – Combined

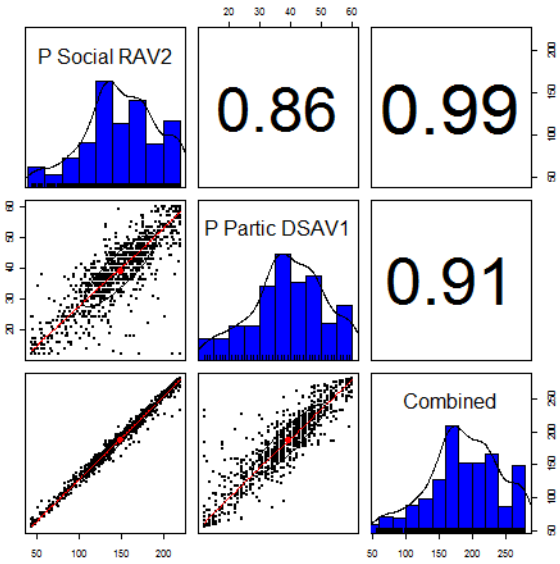


Figure 5.22.4: Scatter Plot Matrix of Raw Summed Scores

5.22.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.22.1: summarizes the results. For P Social RAV2, Cronbach's alpha internal consistency reliability estimate was 0.988 and adjusted (corrected for overlap) item-total correlations ranged from 0.65 to 0.846. For P Partic DSAV1, alpha was 0.957 and adjusted item-total correlations ranged from 0.685 to 0.834. For the 56 items, alpha was 0.99 and adjusted item-total correlations ranged from 0.633 to 0.847.

Table 5.22.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
P Social RAV2	44	0.988	0.650	0.805	0.846
P Partic DSAV1	12	0.957	0.685	0.787	0.834
Combined	56	0.990	0.633	0.789	0.847

5.22.3. 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 5.22.2 summarizes the model fit statistics. For P Social RAV2, the fit statistics were as follows: CFI = 0.941, TLI = 0.939, and RMSEA = 0.1. For P Partic DSAV1, CFI = 0.975, TLI = 0.97, and RMSEA = 0.13. For the 56 items, CFI = 0.928, TLI = 0.925, and RMSEA = 0.093. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.22.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
P Social RAV2	44	1010	0.941	0.939	0.100
P Partic DSAV1	12	1010	0.975	0.970	0.130
Combined	56	1010	0.928	0.925	0.093

5.22.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 56 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 56 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 44 P Social RAV2 items served as anchor items to transform the item parameter estimates for the P Partic DSAV1 items onto the P Social RAV2 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 5.23.3 IRT Linking Constants shows the additive (A) and multiplicative (B) transformation constants derived from the four linking methods. For fixed-parameter calibration, the item parameters for the P Social RAV2 items were constrained to their final bank values, while the P Partic DSAV1 items were calibrated, under the constraints imposed by the anchor items.

Table 5.22.3: IRT Linking Constants

	A	B
Mean/Mean	0.935	-0.360
Mean/Sigma	0.951	-0.357
Haebara	0.934	-0.347
Stocking-Lord	0.946	-0.355

The item parameter estimates for the P Partic DSAV1 items were linked to the P Social RAV2 metric using the transformation constants shown in Table 5.22.3. The P Partic DSAV1 item parameter estimates from the fixed-parameter calibration are considered already on the P Social RAV2 metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for P Partic DSAV1 as shown in Figure 5.22.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.22.6 displays the differences on the vertical axis.

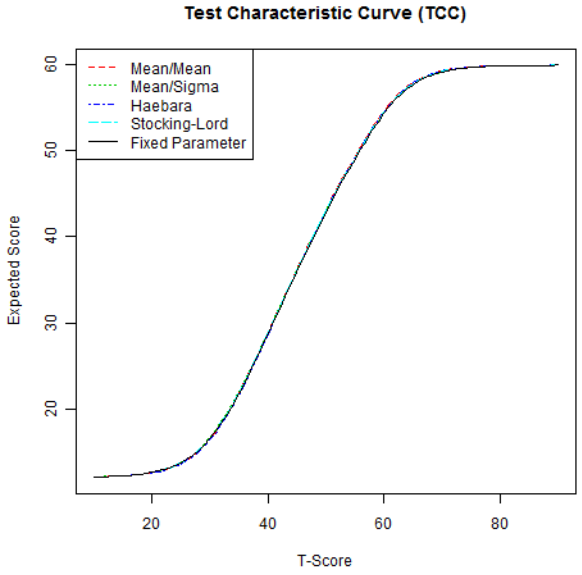


Figure 5.22.5: Test Characteristic Curves (TCC) from Different Linking Methods

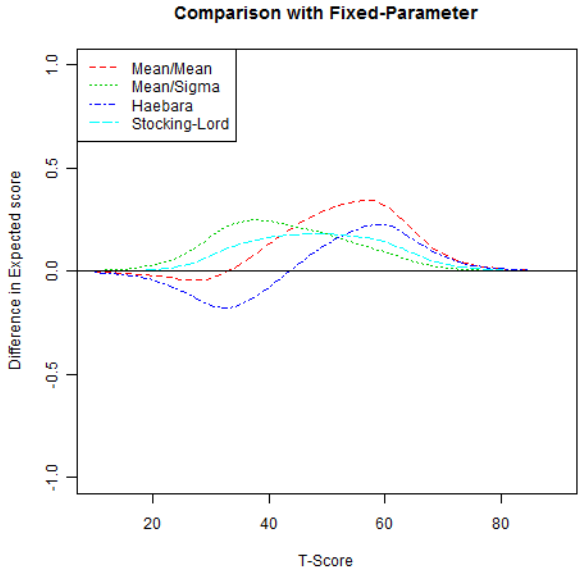


Figure 5.22.6: Difference in Test Characteristic Curves (TCC)

Table 5.22.4: Fixed-Parameter Estimates shows the fixed-parameter calibration item parameter estimates for P Partic DSAV1. The marginal reliability estimate for P Partic DSAV1 based on the item parameter estimates was 0.942. The marginal reliability estimates for P Social RAV2 and the combined set were 0.977 and 0.984, respectively. The slope parameter estimates for P Partic DSAV1 ranged from 1.84 to 2.98 with a mean of 2.58. The slope parameter estimates for P Social RAV2 ranged from 2.12 to 4.75 with a mean of 3.53. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.22.7 displays the scale information functions for P Social RAV2, P Partic DSAV1, and the combined set of 56. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.22.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.22.4: Fixed-Parameter Estimates for PROMIS Satisfaction with Participation in Discretionary Social Activities

a	cb1	cb2	cb3	cb4	NCAT
1.837	-1.805	-0.923	0.155	1.254	5
2.422	-1.564	-0.836	0.046	1.084	5
2.548	-1.591	-0.826	-0.072	0.883	5
2.822	-1.665	-0.924	-0.164	0.780	5
2.644	-1.685	-0.968	-0.165	0.786	5
2.837	-1.549	-0.846	-0.061	0.839	5
2.889	-1.461	-0.833	-0.071	0.757	5
2.982	-1.678	-0.947	-0.212	0.714	5
2.548	-1.727	-0.962	-0.126	0.798	5
2.548	-1.496	-0.846	0.042	0.957	5
2.099	-2.122	-1.253	-0.379	0.609	5
2.765	-1.646	-0.935	-0.155	0.687	5

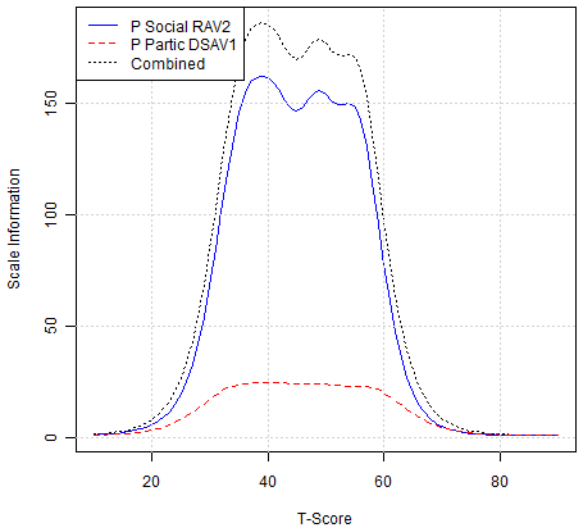


Figure 5.22.7: Comparison of Scale Information Functions

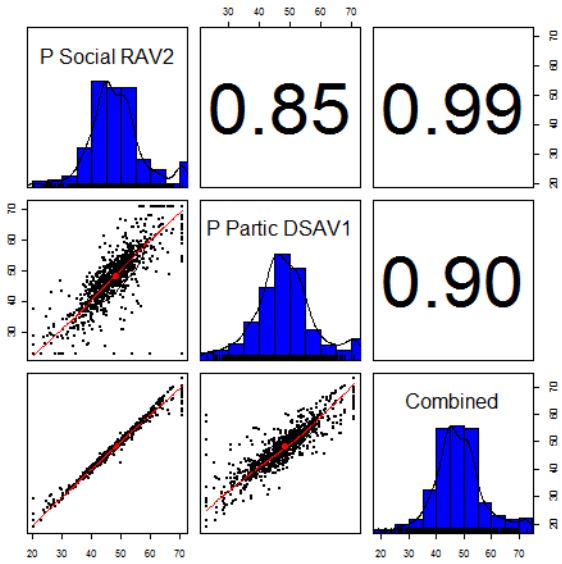


Figure 5.22.8: Comparison of IRT Scaled Scores

5.22.5. 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 P Partic DSAV1 to a scaled score on P Social RAV2 can be useful. Based on the P Partic DSAV1 item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 60 can be used to map simple raw summed scores from P Partic DSAV1 to T-score values linked to the P Social RAV2 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.

5.22.6. Equipercentile Linking

We mapped each raw summed score point on P Partic DSAV1 to a correspond- ing scaled score on P Social RAV2 by identifying scores on P Social RAV2 that have the same percentile ranks as scores on P Partic DSAV1. 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 5.22.9 displays the cumulative distribution functions of the measures. Figure 5.22.10 shows the equipercentile linking functions based on raw summed scores, from P Partic DSAV1 to P Social RAV2. 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 61 and 62 show the equipercentile crosswalk tables. The result shown in Appendix Table 61 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 62 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).

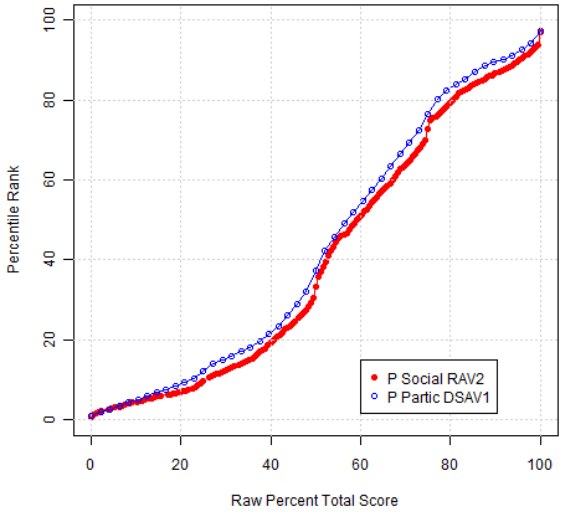


Figure 5.22.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

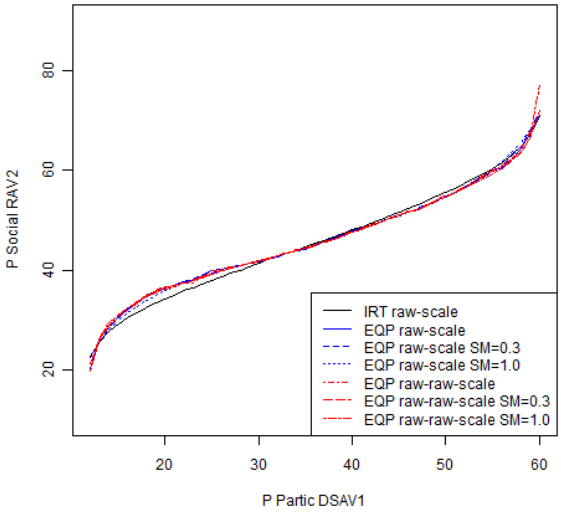


Figure 5.22.10: Equipercentile Linking Functions

5.22.7. 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 5.22.10, 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 5.22.5 reports four statistics summarizing the current sample in terms of the differences between the P Social RAV2 T-scores and P Partic DSAV1 scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.22.10), 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 (0.854), followed by EQP raw-scale SM=1.0 (0.853). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). EQP raw-scale SM=1.0 yielded smallest RMSD (5.221), followed by EQP raw- raw-scale SM=0.0 (5.257)

Table 5.22.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.854	0.005	5.324	5.321
IRT raw-scale	0.853	-0.015	5.309	5.307
EQP raw-scale SM=0.0	0.849	0.098	5.290	5.289
EQP raw-scale SM=0.3	0.851	0.013	5.290	5.287
EQP raw-scale SM=1.0	0.853	0.049	5.223	5.221
EQP raw-raw-scale SM=0.0	0.850	0.075	5.259	5.257
EQP raw-raw-scale SM=0.3	0.850	0.069	5.272	5.270
EQP raw-raw-scale SM=1.0	0.842	-0.139	5.660	5.659

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1007) over a large number of replications (i.e., 10,000).

Table 5.22.6: Comparison of Resampling Results 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 P Social RAV2 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-scale SM=1.0 produced the smallest standard error, 0.582. That is, the difference

between the mean P Social RAV2 T-score and the mean equated P Partic DSAV1 T-score based on a similar sample of 75 cases is expected to be around ±1.16 (i.e., 2 × 0.582).

Table 5.22.6: Comparison of Resampling Results[1]

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	-0.005	1.045	0.007	0.723	0.012	0.592
IRT raw-scale	-0.015	1.058	-0.019	0.745	-0.007	0.593
EQP raw-scale SM=0.0	0.094	1.039	0.094	0.741	0.092	0.585
EQP raw-scale SM=0.3	0.010	1.025	0.014	0.728	0.011	0.590
EQP raw-scale SM=1.0	0.058	1.027	0.047	0.709	0.054	0.582
EQP raw-raw-scale SM=0.0	0.098	1.057	0.085	0.728	0.072	0.583
EQP raw-raw-scale SM=0.3	0.075	1.044	0.064	0.732	0.068	0.583
EQP raw-raw-scale SM=1.0	-0.142	1.127	-0.123	0.785	-0.149	0.625

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

5.23. PROMIS Satisfaction with Social Roles and Activities (v2.0) and PROMIS Satisfaction with Participation in Social Roles (v1.0)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Social Functioning, namely the PROMIS Satisfaction with Social Roles and Activities v2.0 (P Social RAV2) item bank (44 items) and the PROMIS Satisfaction with Participation in Social Roles v1.0 (P Partic SRV1) item bank (14 items). Both instruments were scaled such that higher scores represent higher levels of Social. We did not exclude any participants because of missing responses, leaving a final sample of N=1006. 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.

5.23.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 220 for P Social RAV2 and 70 for P Partic SRV1. Figures 5.23.1 and 5.23.2 graphically display the raw summed score distributions of the two measures. Figure 5.23.3 shows the distribution for the combined. Figure 5.23.4 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 P Social RAV2 and P Partic SRV1 was 0.9. The disattenuated (corrected for unreliabilities) correlation between P Social RAV2 and P Partic SRV1 was 0.92. The correlations between the combined score and the measures were 0.99 and 0.94 for P Social RAV2 and P Partic SRV1, respectively.

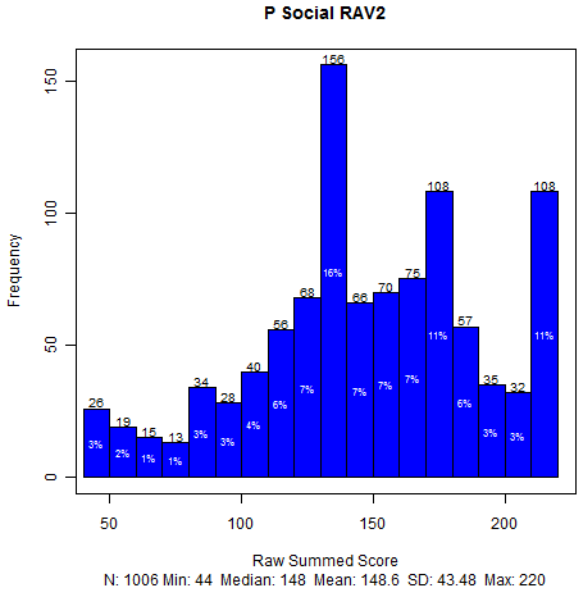


Figure 5.23.1: Raw Summed Score Distribution – PROMIS Satisfaction with Social Roles and Activities v2.0

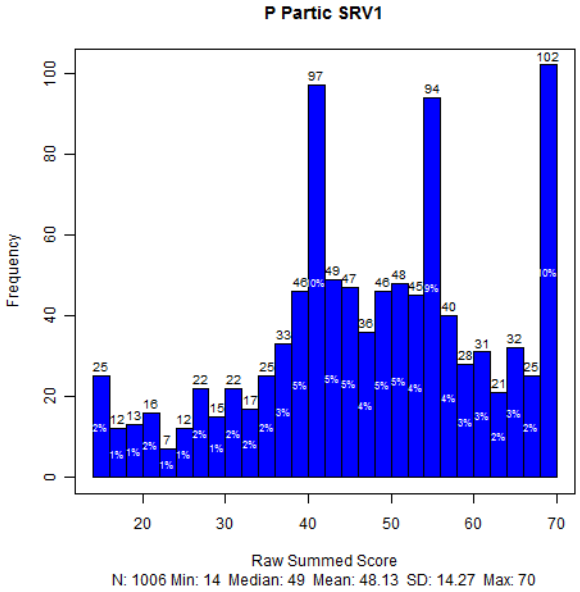


Figure 5.23.2: Raw Summed Score Distribution – Satisfaction with Participation in Social Roles v1.0

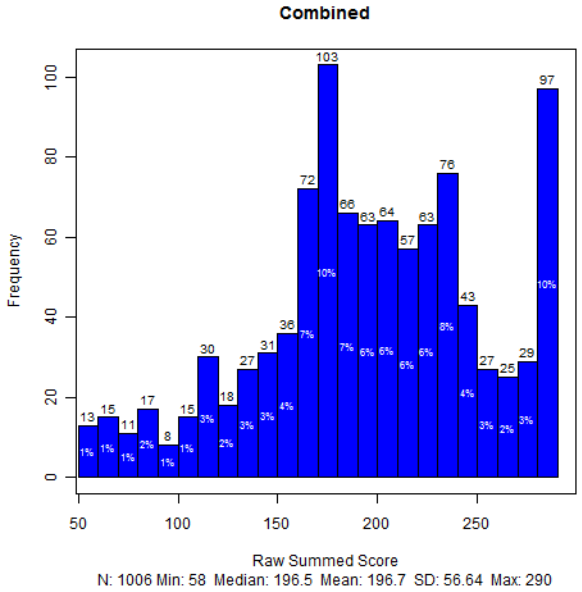


Figure 5.23.3: Raw Summed Score Distribution – Combined

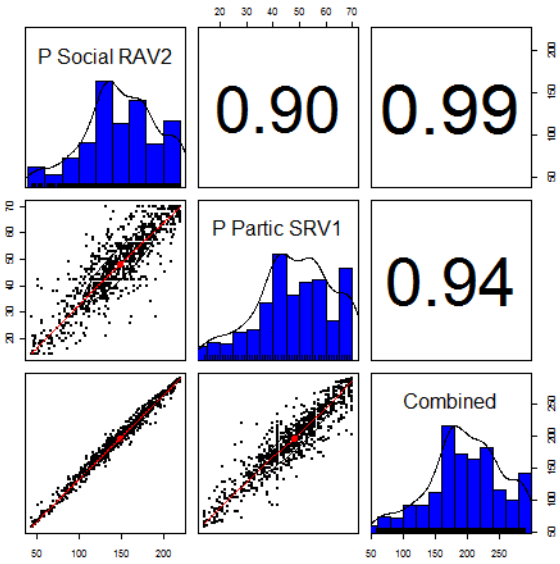


Figure 5.23.4: Scatter Plot Matrix of Raw Summed Scores

5.23.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.23.1 summarizes the results. For P Social RAV2, Cronbach's alpha internal consistency reliability estimate was 0.988 and adjusted (corrected for overlap) item-total correlations ranged from 0.65 to 0.846. For P Partic SRV1, alpha was 0.969 and adjusted item-total correlations ranged from 0.789 to 0.841. For the 58 items, alpha was 0.991 and adjusted item-total correlations ranged from 0.642 to 0.836.

Table 5.23.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
P Social RAV2	44	0.988	0.650	0.805	0.846
P Partic SRV1	14	0.969	0.789	0.815	0.841
Combined	58	0.991	0.642	0.798	0.836

5.23.3. 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 5.23.2 summarizes the model fit statistics. For P Social RAV2, the fit statistics were as follows: CFI = 0.941, TLI = 0.939, and RMSEA = 0.1. For P Partic SRV1, CFI = 0.971, TLI = 0.966, and RMSEA = 0.141. For the 58 items, CFI = 0.93, TLI = 0.928, and RMSEA = 0.091. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.23.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
P Social RAV2	44	1010	0.941	0.939	0.100
P Partic SRV1	14	1010	0.971	0.966	0.141
Combined	58	1010	0.930	0.928	0.091

5.23.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 58 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 58 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 44 P Social RAV2 items served as anchor items to transform the item parameter estimates for the P Partic SRV1 items onto the P Social RAV2 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 5.23.3:IRT Linking Constants shows the additive (A) and multiplicative (B) transformation constants derived from the four linking methods. For fixed-parameter calibration, the item parameters for the P Social RAV2 items were constrained to their final bank values, while the P Partic SRV1 items were calibrated, under the constraints imposed by the anchor items.

Table 5.23.3: IRT Linking Constants

	A	B
Mean/Mean	0.935	-0.381
Mean/Sigma	0.958	-0.377
Haebara	0.939	-0.367
Stocking-Lord	0.951	-0.373

The item parameter estimates for the P Partic SRV1 items were linked to the P Social RAV2 metric using the transformation constants shown in Table 5.23.3. The P Partic SRV1 item parameter estimates from the fixed-parameter calibration are considered already on the P Social RAV2 metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for P Partic SRV1 as shown in Figure 5.23.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.23.6 displays the differences on the vertical axis.

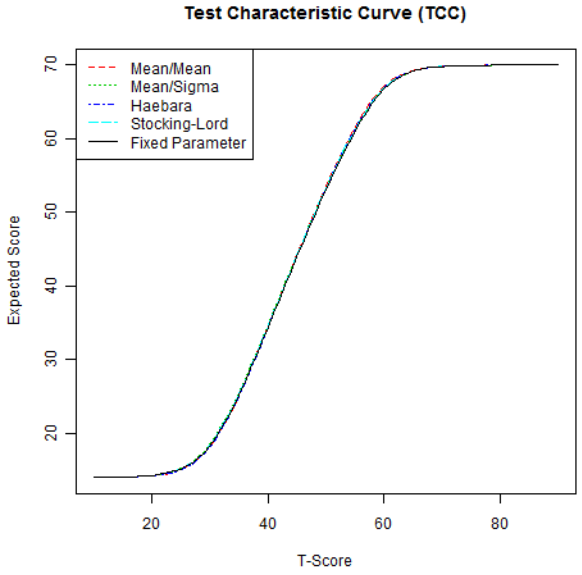


Figure 5.23.5: Test Characteristic Curves (TCC) from Different Linking Methods

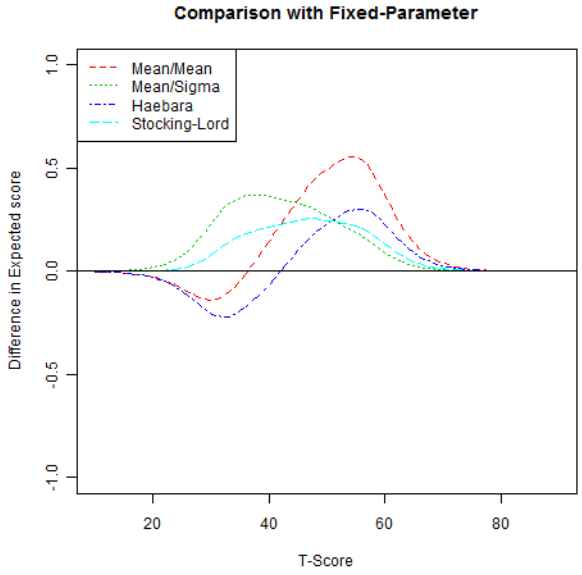


Figure 5.23.6: Difference in Test Characteristic Curves (TCC)

Table 5.23.4: Fixed-Parameter for shows the fixed-parameter calibration item parameter estimates for P Partic SRV1. The marginal reliability estimate for P Partic SRV1 based on the item parameter estimates was 0.944. The marginal reliability estimates for P Social RAV2 and the combined set were 0.977 and 0.981, respectively. The slope parameter estimates for P Partic SRV1 ranged from 2.84 to 3.41 with a mean of 3.06. The slope parameter estimates for P Social RAV2 ranged from 2.12 to 4.75 with a mean of 3.53. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.23.7 displays the scale information functions for P Social RAV2, P Partic SRV1, and the combined set of 58. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.23.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.23.4: Fixed-Parameter Estimates for PROMIS Satisf w/ Partic in Social Roles

a	cb1	cb2	cb3	cb4	NCAT
2.966	-1.597	-0.974	-0.248	0.564	5
2.855	-1.664	-0.986	-0.250	0.610	5
3.412	-1.706	-1.041	-0.316	0.452	5
3.193	-1.652	-0.944	-0.254	0.543	5
2.991	-1.517	-0.912	-0.092	0.728	5
3.183	-1.653	-0.988	-0.229	0.492	5
3.271	-1.540	-0.952	-0.268	0.530	5
2.844	-1.612	-0.935	-0.245	0.589	5
2.893	-1.645	-0.880	-0.124	0.773	5
2.919	-1.650	-0.920	-0.234	0.564	5
3.059	-1.776	-0.970	-0.253	0.541	5
2.994	-1.742	-0.996	-0.293	0.473	5
3.356	-1.682	-0.964	-0.337	0.461	5
2.947	-1.709	-1.081	-0.323	0.392	5

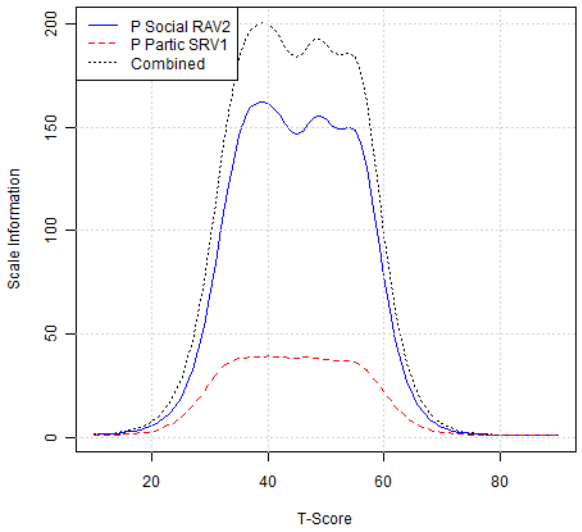


Figure 5.23.7: Comparison of Scale Information Functions

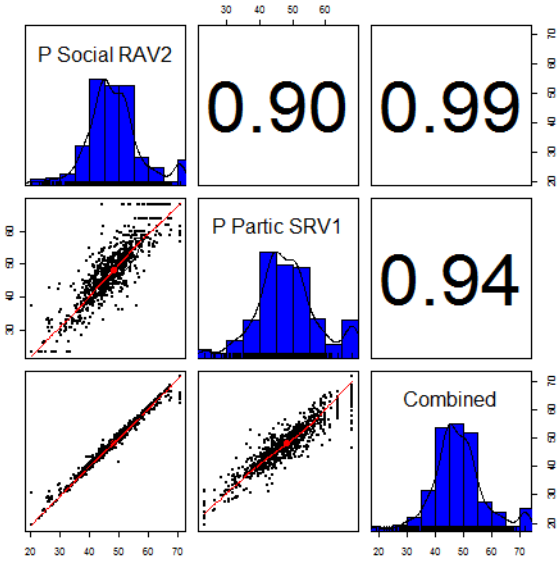


Figure 5.23.8: Comparison of IRT Scaled Scores

5.23.5. 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 P Partic SRV1 to a scaled score on P Social RAV2 can be useful. Based on the P Partic SRV1 item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 63 can be used to map simple raw summed scores from P Partic SRV1 to T-score values linked to the P Social RAV2 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.

5.23.6. Equipercentile Linking

We mapped each raw summed score point on P Partic SRV1 to a corresponding scaled score on P Social RAV2 by identifying scores on P Social RAV2 that have the same percentile ranks as scores on P Partic SRV1. 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 5.23.9 displays the cumulative distribution functions of the measures. Figure 5.23.10 shows the equipercentile linking functions based on raw summed scores, from Partic

SRV1 to P Social RAV2. 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 64 and 65 show the equipercentile crosswalk tables. The result shown in Appendix Table 64 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 65 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).

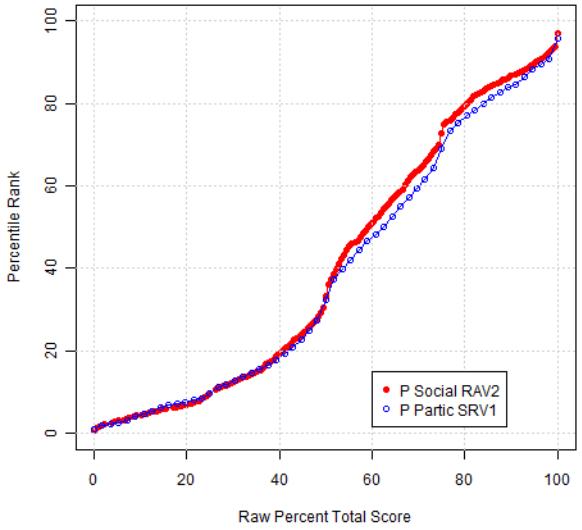


Figure 5.23.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

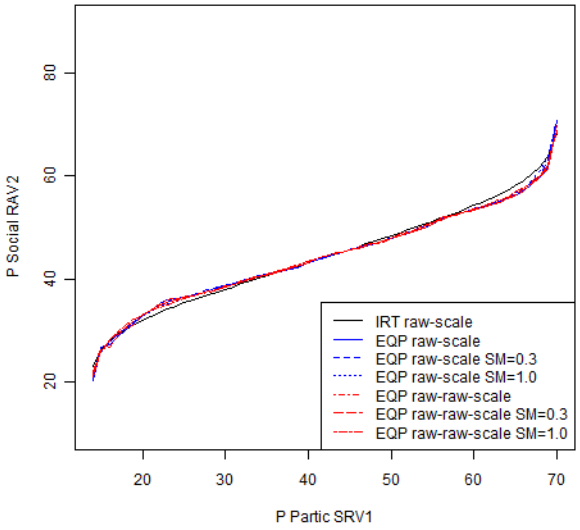


Figure 5.23.10: Equipercentile Linking Functions

5.23.7. 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 5.23.10, 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 5.23.5 reports four statistics summarizing the current sample in terms of the differences between the P Social RAV2 T-scores and P Partic SRV1 scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.23.10), 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=1.0 produced the best result (0.901), followed by EQP raw-raw-scale SM=0.3 (0.9). 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 (4.256), followed by EQP raw-raw-scale SM=0.3 (4.273).

Table 5.23.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.900	-0.129	4.324	4.324
IRT raw-scale	0.900	-0.163	4.328	4.329
EQP raw-scale SM=0.0	0.899	-0.062	4.412	4.410
EQP raw-scale SM=0.3	0.899	-0.116	4.406	4.405
EQP raw-scale SM=1.0	0.899	-0.176	4.400	4.401
EQP raw-raw-scale SM=0.0	0.900	0.016	4.317	4.315
EQP raw-raw-scale SM=0.3	0.900	0.040	4.275	4.273
EQP raw-raw-scale SM=1.0	0.901	0.037	4.258	4.256

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1006) over a large number of replications (i.e., 10,000).

Table 5.23.6: Comparison of Resampling Results 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

P Social RAV2 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, 0.469. That is, the difference between the mean P Social RAV2 T-score and the mean equated P Partic SRV1 T-score based on a similar sample of 75 cases is expected to be around ±0.94 (i.e., 2 × 0.469).

Table 5.23.6: Comparison of Resampling Results[2]

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	-0.136	0.857	-0.122	0.597	-0.123	0.477
IRT raw-scale	-0.171	0.857	-0.169	0.594	-0.162	0.479
EQP raw-scale SM=0.0	-0.054	0.870	-0.061	0.602	-0.062	0.489
EQP raw-scale SM=0.3	-0.107	0.869	-0.118	0.610	-0.114	0.488
EQP raw-scale SM=1.0	-0.169	0.870	-0.173	0.604	-0.178	0.486
EQP raw-raw-scale SM=0.0	0.014	0.857	0.019	0.597	0.016	0.483
EQP raw-raw-scale SM=0.3	0.039	0.836	0.047	0.589	0.043	0.475
EQP raw-raw-scale SM=1.0	0.044	0.832	0.035	0.590	0.035	0.469

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

5.24 Neuro-QOL Positive Affect & Well-being and NIH Toolbox Life Satisfaction

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Emotion, namely the Neuro-QOL Positive Affect & Well-being item bank (23 items) and NIH Toolbox Life Satisfaction (10 items). Neuro-QOL Positive Affect & Well-being was scaled such that higher scores represent higher levels of Emotion. 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. Our sample consisted of 1,016 participants (N = 1,015 participants with complete responses).

5.24.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 115 for Neuro-QOL Pos Affect and 60 for TB Life Satisfaction. Figures 5.24.1 and 5.24.2 graphically display the raw summed score distributions of the two measures. Figure 5.24.3 shows the distribution for the combined. Figure 5.24.4 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 Neuro-QOL Pos Affect and TB Life Satisfaction was 0.77. The disattenuated (corrected for unreliabilities) correlation between Neuro-QOL Pos Affect and TB Life Satisfaction was 0.81. The correlations between the combined score and the measures were 0.97 and 0.91 for Neuro-QOL Pos Affect and TB Life Satisfaction, respectively.

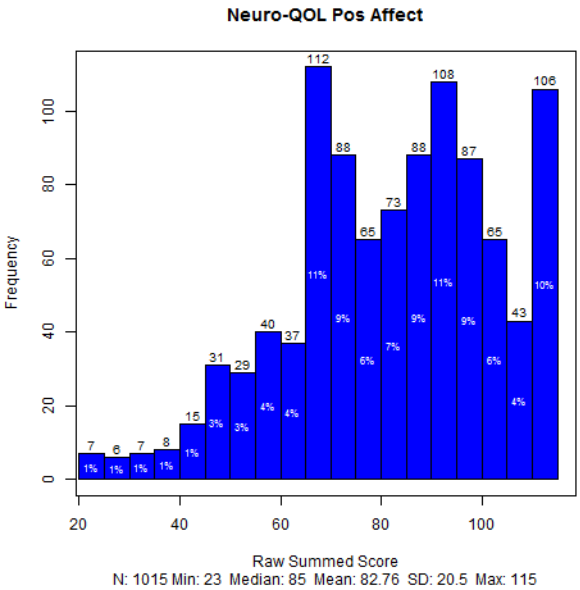


Figure 5.24.1: Raw Summed Score Distribution – Neuro-QOL Positive Affect & Well-being

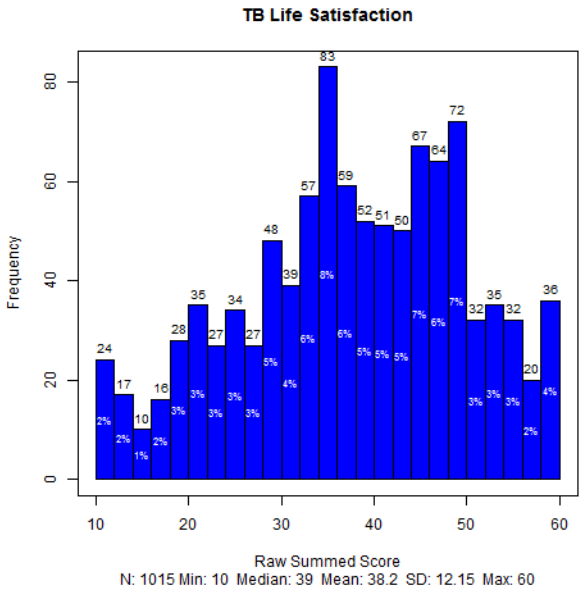


Figure 5.24.2: Raw Summed Score Distribution – NIH Toolbox Life Satisfaction

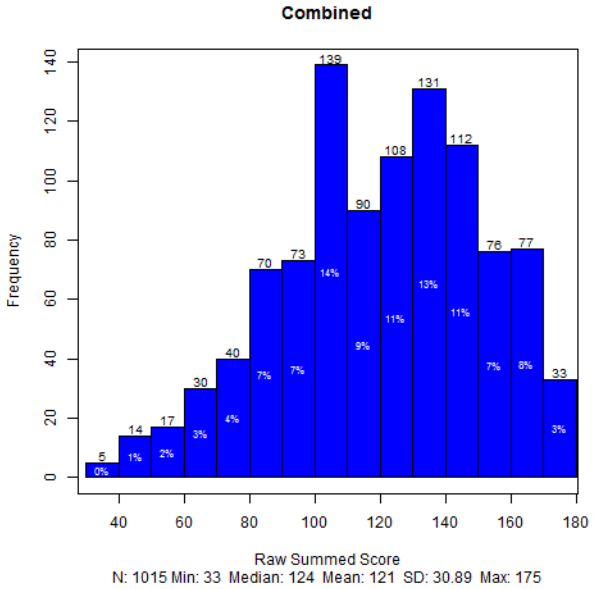


Figure 5.24.3: Raw Summed Score Distribution – Combined

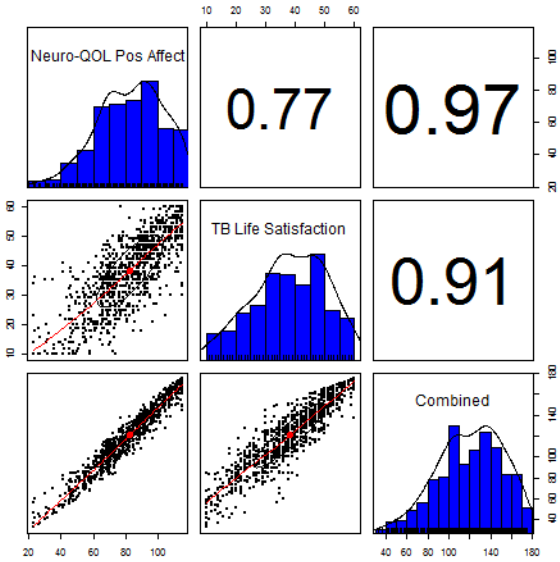


Figure 5.24.4: Scatter Plot Matrix of Raw Summed Scores

5.24.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.24.1 summarizes the results. For Neuro-QOL Pos Affect, Cronbach's alpha internal consistency reliability estimate was 0.977 and adjusted (corrected for overlap) item-total correlations ranged from 0.733 to 0.859. For TB Life Satisfaction, alpha was 0.938 and adjusted item-total correlations ranged from 0.574 to 0.862. For the 33 items, alpha was 0.977 and adjusted item-total correlations ranged from 0.548 to 0.844.

Table 5.24.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
Neuro-QOL Pos Affect	23	0.977	0.733	0.794	0.859
TB Life Satisfaction	10	0.938	0.574	0.770	0.862
Combined	33	0.977	0.548	0.758	0.844

5.24.3. 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 5.24.2 summarizes the model fit statistics. For Neuro-QOL Pos Affect, the fit statistics were as follows: CFI = 0.969, TLI = 0.966, and RMSEA = 0.104. For TB Life Satisfaction, CFI = 0.985, TLI = 0.98, and RMSEA = 0.117. For the 33 items, CFI = 0.92, TLI = 0.914, and RMSEA = 0.128. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.24.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
Neuro-QOL Pos Affect	23	1016	0.969	0.966	0.104
TB Life Satisfaction	10	1016	0.985	0.980	0.117
Combined	33	1016	0.920	0.914	0.128

5.24.4. 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 23 Neuro-QOL Pos Affect items served as anchor items to transform the item parameter estimates for the TB Life Satisfaction items onto the Neuro-QOL Pos Affect 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 5.24.3 shows the additive (A) and multiplicative (B) transformation constants derived from the four linking methods. For fixed-parameter calibration, the item parameters for the Neuro-QOL Pos Affect items were constrained to their final bank values, while the TB Life Satisfaction items were calibrated, under the constraints imposed by the anchor items.

Table 5.24.3: IRT Linking Constants

	A	B
Mean/Mean	0.767	0.045
Mean/Sigma	0.924	0.124
Haebara	0.906	0.116
Stocking-Lord	0.903	0.113

The item parameter estimates for the TB Life Satisfaction items were linked to the Neuro-QOL Pos Affect metric using the transformation constants shown in Table 5.24.3. The TB Life Satisfaction item parameter estimates from the fixed- parameter calibration are considered already on the Neuro-QOL Pos Affect metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for TB Life Satisfaction as shown in Figure 5.24.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.24.6 displays the differences on the vertical axis.

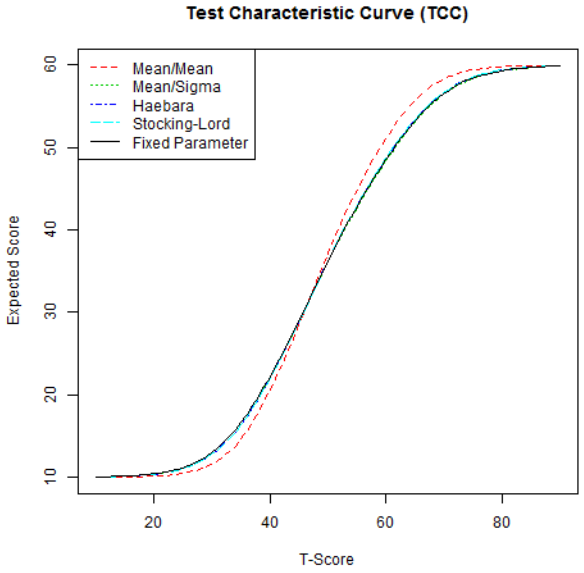


Figure 5.24.5: Test Characteristic Curves (TCC) from Different Linking Methods

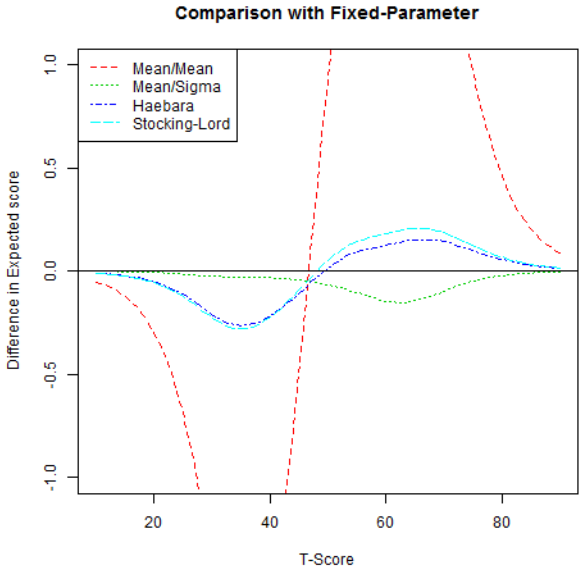


Figure 5.24.6: Difference in Test Characteristic Curves (TCC)

Table 5.23.4: Fixed-Parameter Estimates for shows the fixed-parameter calibration item parameter estimates for TB Life Satisfaction. The marginal reliability estimate for TB Life Satisfaction based on the item parameter estimates was 0.932. The marginal reliability estimates for Neuro-QOL Pos Affect and the combined set were 0.977 and 0.985, respectively. The slope parameter estimates for TB Life Satisfaction ranged from 1.5 to 2.85 with a mean of 2.27. The slope parameter estimates for Neuro-QOL Pos Affect ranged from 2.66 to 6.61 with a mean of 4.02. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.24.7 displays the scale information functions for Neuro-QOL Pos Affect, TB Life Satisfaction, and the combined set of 33. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.24.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.24.4: Fixed-Parameter Estimates for NIH Toolbox Life Satisfaction

a	cb1	cb2	cb3	cb4	cb5	cb6	NCAT
1.504	-1.650	-0.369	0.407	1.481			5
2.288	-1.195	-0.454	0.055	0.504	1.020	1.990	7
1.701	-1.349	-0.505	0.059	0.555	1.133	2.179	7
2.776	-1.400	-0.867	-0.443	-0.035	0.385	1.360	7
2.229	-1.537	-0.896	-0.451	-0.017	0.496	1.486	7
2.381	-1.359	-0.746	-0.292	0.174	0.678	1.615	7
2.442	-1.557	-0.716	-0.106	1.053			5
2.286	-1.471	-0.491	0.343	1.522			5
2.847	-1.574	-0.841	-0.144	1.033			5
2.235	-1.495	-0.511	0.253	1.524			5

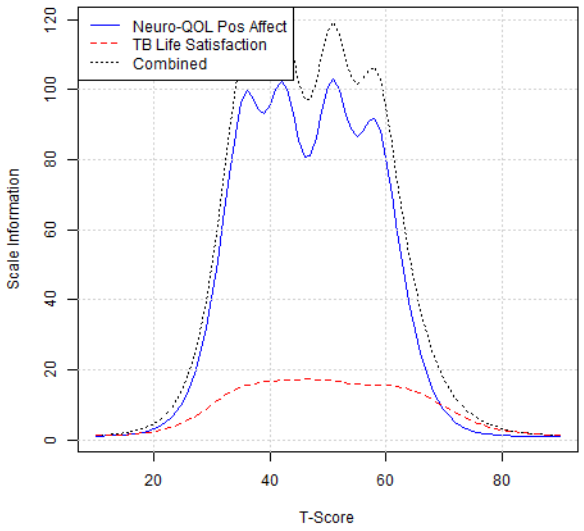


Figure 5.24.7: Comparison of Scale Information Functions

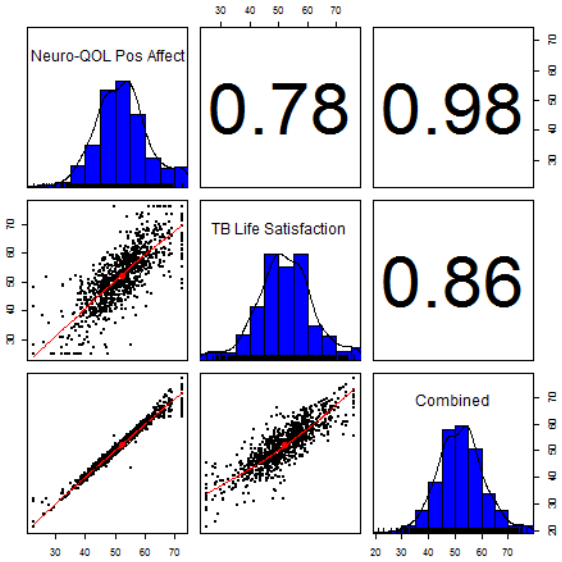


Figure 5.24.8: Comparison of IRT Scaled Scores

5.24.5. 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 TB Life Satisfaction to a scaled score on Neuro-QOL Pos Affect can be useful. Based on the TB Life Satisfaction item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 66 can be used to map simple raw summed scores from TB Life Satisfaction to T-score values linked to the Neuro-QOL Pos Affect metric. Each raw summed score point and corresponding Neuro-QOL 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.

5.24.6. Equipercentile Linking

We mapped each raw summed score point on TB Life Satisfaction to a corresponding scaled score on Neuro-QOL Pos Affect by identifying scores on Neuro- QOL Pos Affect that have the same percentile ranks as scores on TB Life Satisfaction. 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 5.24.9 displays the cumulative distribution functions of the measures. Figure 5.24.10 shows the equipercentile linking functions based on raw summed

scores, from TB Life Satisfaction to Neuro-QOL Pos Affect. 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 67 and 68 show the equipercentile crosswalk tables. The result shown in Appendix Table 67 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 68 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).

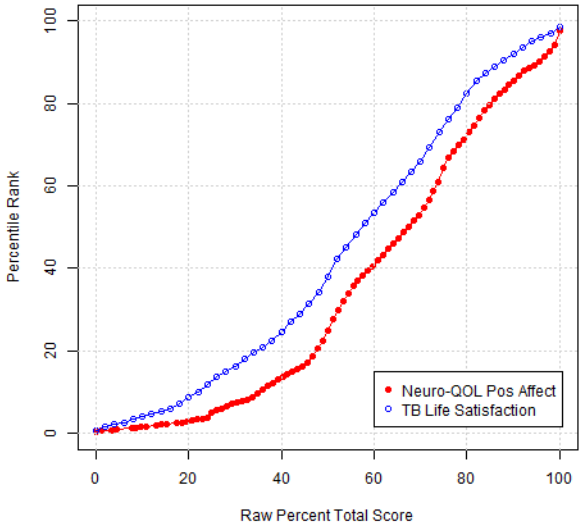


Figure 5.24.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

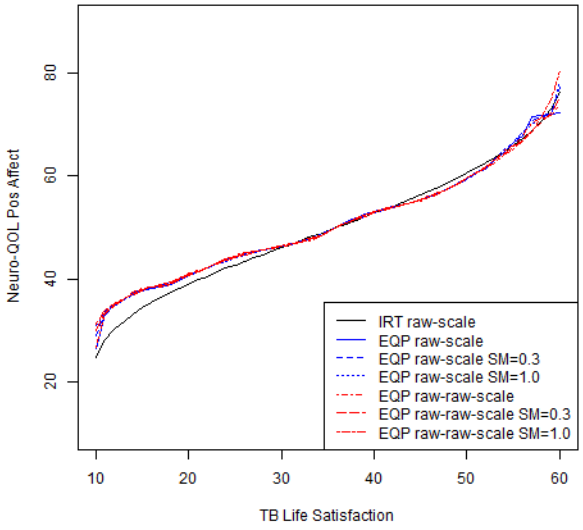


Figure 5.24.10: Equipercentile Linking Functions

5.24.7. 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 5.24.10, 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 5.24.5 reports four statistics summarizing the current sample in terms of the differences between the Neuro-QOL Pos Affect T-scores and TB Life Satisfaction scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.24.10), 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 (0.779), followed by IRT raw-scale (0.768). Similar results were found in terms of the standard de- viation of differences and root mean squared difference (RMSD). EQP raw-raw-scale SM=0.3 yielded smallest RMSD (6.092), followed by EQP raw-raw-scale SM=0.0 (6.126).

Table 5.24.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.779	0.139	6.379	6.377
IRT raw-scale	0.768	0.137	6.502	6.501
EQP raw-scale SM=0.0	0.766	0.046	6.146	6.143
EQP raw-scale SM=0.3	0.766	-0.109	6.216	6.214
EQP raw-scale SM=1.0	0.766	-0.173	6.230	6.230
EQP raw-raw-scale SM=0.0	0.767	0.038	6.129	6.126
EQP raw-raw-scale SM=0.3	0.768	-0.008	6.095	6.092
EQP raw-raw-scale SM=1.0	0.763	-0.150	6.282	6.281

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1015) over a large number of replications (i.e., 10,000).

Table 5.24.6: Comparison of Resampling Results 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

Neuro-QOL Pos Affect 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, 0.678. That is, the difference between the mean Neuro-QOL Pos Affect T-score and the mean equated TB Life Satisfaction T-score based on a similar sample of 75 cases is expected to be around ±1.36 (i.e., 2 × 0.678).

Table 5.24.6: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	0.123	1.253	0.129	0.881	0.135	0.704
IRT raw-scale	0.126	1.280	0.136	0.899	0.144	0.725
EQP raw-scale SM=0.0	0.052	1.227	0.041	0.849	0.042	0.680
EQP raw-scale SM=0.3	-0.121	1.225	-0.104	0.859	-0.104	0.690
EQP raw-scale SM=1.0	-0.161	1.213	-0.177	0.872	-0.173	0.701
EQP raw-raw-scale SM=0.0	0.056	1.215	0.032	0.833	0.041	0.684
EQP raw-raw-scale SM=0.3	0.024	1.205	-0.018	0.846	-0.010	0.678
EQP raw-raw-scale SM=1.0	-0.135	1.253	-0.153	0.865	-0.151	0.700

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

5.25 Neuro-QOL Positive Affect & Well-being and NIH Toolbox Meaning

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Emotion, namely the Neuro-QOL Positive Affect & Well-being item bank (23 items) and NIH Toolbox Meaning (18 items). Neuro-QOL Positive Affect & Well-being was scaled such that higher scores represent higher levels of Emotion. 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. Our sample consisted of 1,016 participants (N = 1,014 for participants with complete responses).

5.25.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 115 for Neuro-QOL Pos Affect and 90 for TB Meaning. Figures 5.25.1 and 5.25.2 graphically display the raw summed score distributions of the two measures. Figure 5.25.3 shows the distribution for the combined. Figure 5.25.4 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 Neuro-QOL Pos Affect and TB Meaning was 0.81. The disattenuated (corrected for unreliabilities) correlation between Neuro-QOL Pos Affect and TB Meaning was 0.84. The correlations between the combined score and the measures were 0.96 and 0.94 for Neuro-QOL Pos Affect and TB Meaning, respectively.

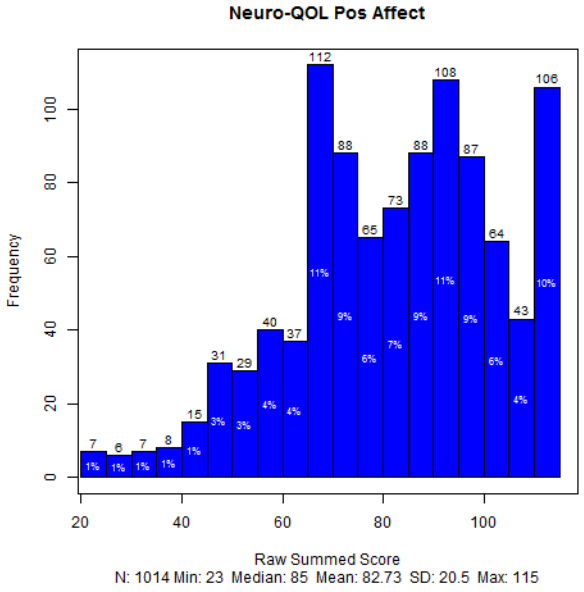


Figure 5.25.1: Raw Summed Score Distribution – Neuro-QOL Positive Affect & Well-being

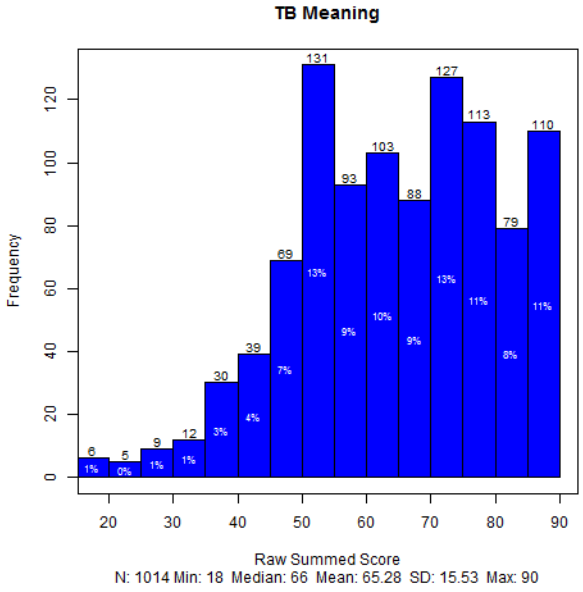


Figure 5.25.2: Raw Summed Score Distribution – NIH Toolbox Meaning

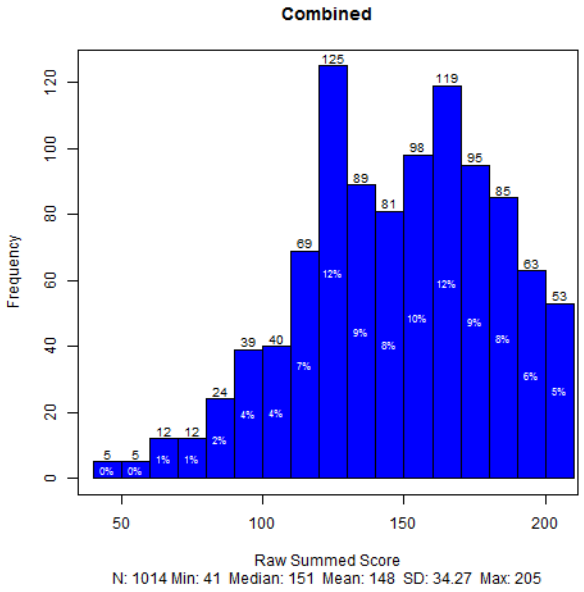


Figure 5.25.3: Raw Summed Score Distribution – Combined

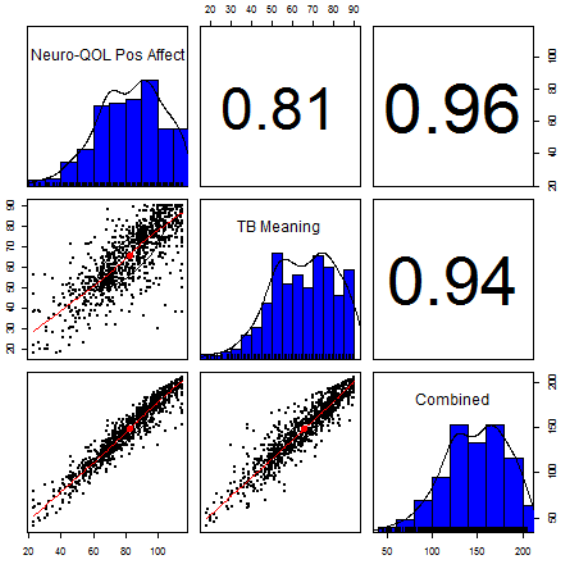


Figure 5.25.4: Scatter Plot Matrix of Raw Summed Scores

5.25.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.25.1 summarizes the results. For Neuro-QOL Pos Affect, Cron- bach's alpha internal consistency reliability estimate was 0.977 and adjusted (corrected for overlap) item-total correlations ranged from 0.733 to 0.859. For TB Meaning, alpha was 0.952 and adjusted item-total correlations ranged from 0.512 to 0.824. For the 41 items, alpha was 0.98 and adjusted item-total correlations ranged from 0.462 to 0.841.

Table 5.25.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
Neuro-QOL Pos Affect	23	0.977	0.733	0.794	0.859
TB Meaning	18	0.952	0.512	0.707	0.824
Combined	41	0.980	0.462	0.730	0.841

5.25.3. 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 5.25.2 summarizes the model fit statistics. For Neuro-QOL Pos Affect, the fit statistics were as follows: CFI = 0.969, TLI = 0.966, and

RMSEA = 0.104. For TB Meaning, CFI = 0.898, TLI = 0.884, and RMSEA = 0.176. For the 41 items, CFI = 0.892, TLI = 0.887, and RMSEA = 0.122. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.25.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
Neuro-QOL Pos Affect	23	1016	0.969	0.966	0.104
TB Meaning	18	1016	0.898	0.884	0.176
Combined	41	1016	0.892	0.887	0.122

5.25.4. 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 23 Neuro-QOL Pos Affect items served as anchor items to transform the item parameter estimates for the TB Meaning items onto the Neuro-QOL Pos Affect 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 5.25.3: IRT Linking Constants shows the additive (A) and multiplicative (B) transformation constants derived from the four linking methods. For fixed-parameter calibration, the item parameters for the Neuro-QOL Pos Affect items were constrained to their final bank values, while the TB Meaning items were calibrated, under the constraints imposed by the anchor items.

Table 5.25.3: IRT Linking Constants

	A	B
Mean/Mean	0.758	-0.007
Mean/Sigma	0.931	0.070
Haebara	0.910	0.060
Stocking-Lord	0.907	0.061

The item parameter estimates for the TB Meaning items were linked to the Neuro-QOL Pos Affect metric using the transformation constants shown in Table 5.25.3. The TB Meaning item parameter estimates from the fixed-parameter calibration are considered already on the Neuro-QOL Pos Affect metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for TB Meaning as shown in Figure 5.25.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.25.6 displays the differences on the vertical axis.

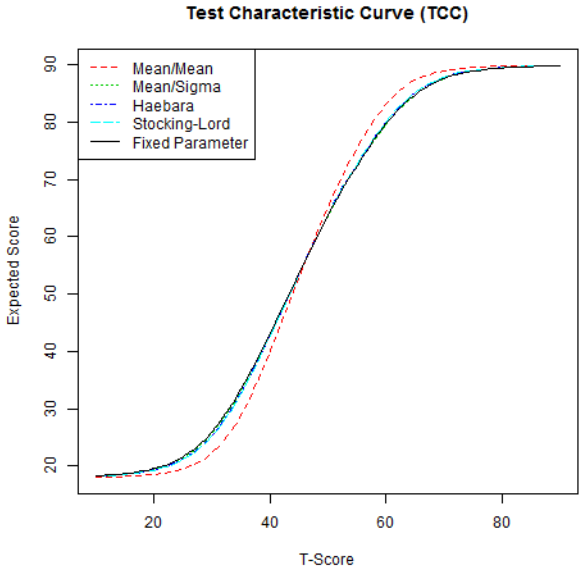


Figure 5.25.5: Test Characteristic Curves (TCC) from Different Linking Methods

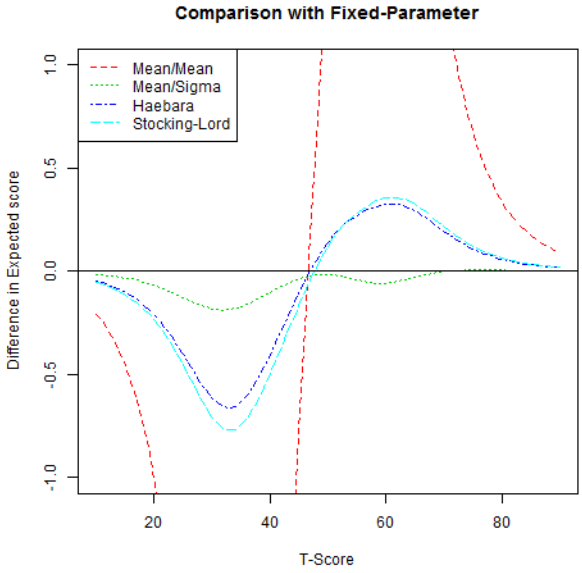


Figure 5.25.6: Difference in Test Characteristic Curves (TCC)

Table 5.25.4: Fixed-Parameter Estimates for shows the fixed-parameter calibration item parameter estimates for TB Meaning. The marginal reliability estimate for TB Meaning based on the item parameter estimates was 0.954. The marginal reliability estimates for Neuro-QOL Pos Affect and the combined set were 0.977 and 0.985, respectively. The slope parameter estimates for TB Meaning ranged from 1.25 to 3.14 with a mean of 2.22. The slope parameter estimates for Neuro-QOL Pos Affect ranged from 2.66 to 6.61 with a mean of 4.02. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.25.7 displays the scale information functions for Neuro-QOL Pos Affect, TB Meaning, and the combined set of 41. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.25.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.25.4: Fixed-Parameter Estimates for NIH Toolbox Meaning

a	cb1	cb2	cb3	cb4	NCAT
1.619	-2.003	-1.056	-0.191	0.746	5
1.725	-1.866	-0.635	-0.001	0.861	5
1.949	-1.744	-0.755	-0.102	0.800	5
1.860	-1.932	-1.023	-0.384	0.561	5
2.190	-1.584	-0.889	-0.241	0.383	5
1.247	-1.960	-0.811	0.384	1.983	5
1.996	-1.548	-0.690	0.189	1.304	5
2.190	-1.655	-0.871	-0.145	1.131	5
2.317	-1.528	-0.693	0.126	1.208	5
2.633	-1.525	-0.843	-0.134	1.067	5
2.240	-1.739	-1.023	-0.376	0.680	5
2.397	-1.502	-0.766	0.014	1.211	5
2.220	-1.764	-0.787	0.022	1.203	5

2.255	-1.730	-1.008	-0.181	1.062	5
2.599	-1.651	-1.133	-0.421	0.577	5
2.605	-1.731	-1.129	-0.412	0.355	5
2.765	-1.553	-0.868	-0.007	0.832	5
3.144	-1.395	-0.678	0.051	0.786	5

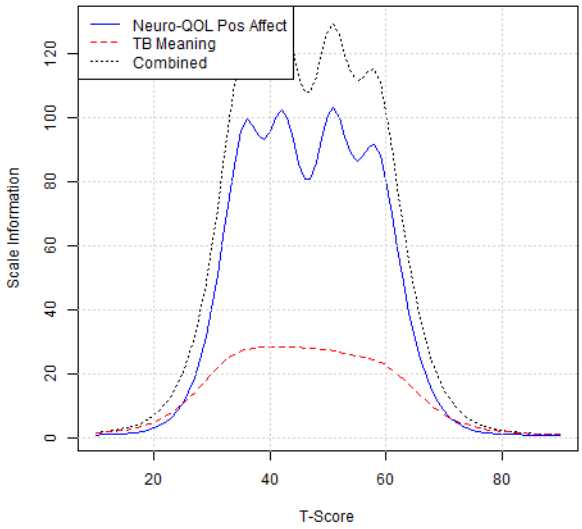


Figure 5.25.7: Comparison of Scale Information Functions

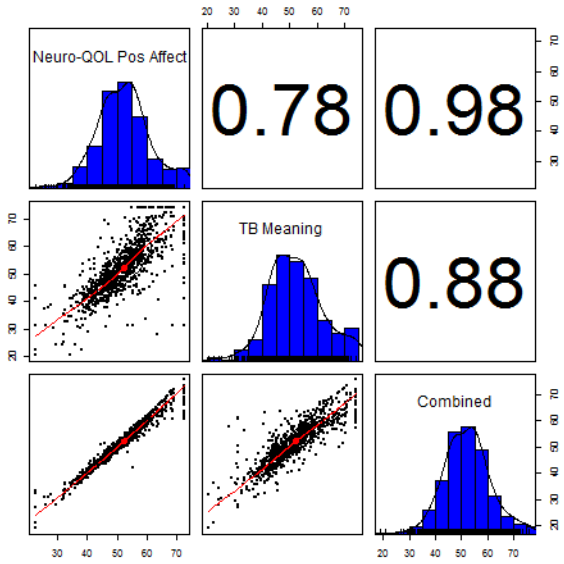


Figure 5.25.8: Comparison of IRT Scaled Scores

5.25.5. 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 TB Meaning to a scaled score on Neuro-QOL Pos Affect can be useful. Based on the TB Meaning item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 69 can be used to map simple raw summed scores from TB Meaning to T-score values linked to the Neuro-QOL Pos Affect metric. Each raw summed score point and corresponding Neuro-QOL 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.

5.25.6. Equipercentile Linking

We mapped each raw summed score point on TB Meaning to a corresponding scaled score on Neuro-QOL Pos Affect by identifying scores on Neuro- QOL Pos Affect that have the same percentile ranks as scores on TB Meaning. 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 5.25.9 displays the cumulative distribution functions of the measures. Figure 5.25.10 shows the equipercentile linking functions based on raw summed scores, from TB Meaning to Neuro-QOL Pos Affect. 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 70 and 71 show the equipercentile crosswalk tables. The result shown in Appendix Table 70 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 71 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).

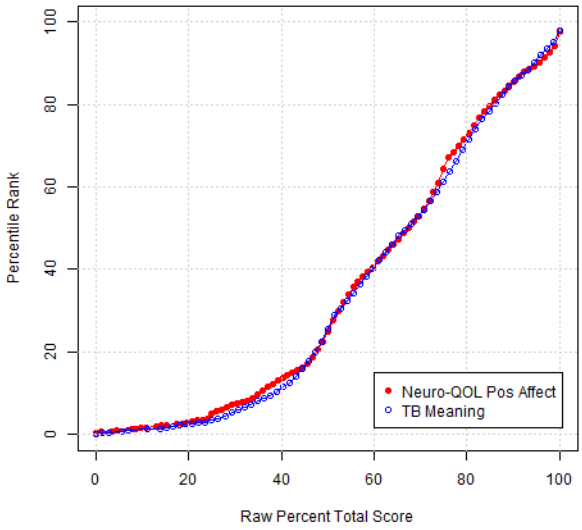


Figure 5.25.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

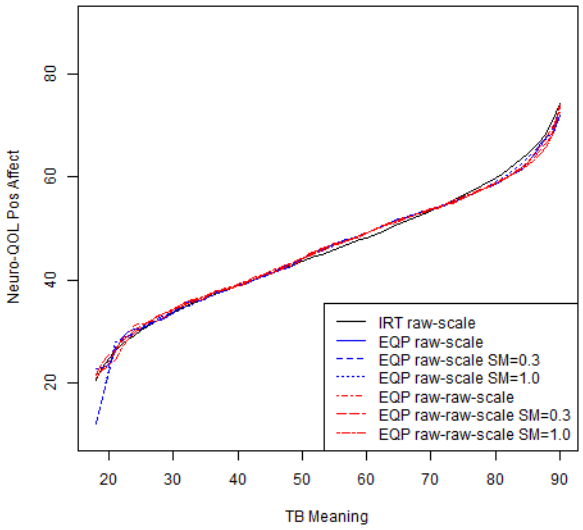


Figure 5.25.10: Equipercentile Linking Functions

5.25.7. 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 5.25.10, 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 5.25.5 reports four statistics summarizing the current sample in terms of the differences between the Neuro-QOL Pos Affect T-scores and TB Meaning scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.25.10), 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 (0.784), followed by IRT raw-scale (0.77). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). EQP raw- scale SM=0.0 yielded smallest RMSD (6.075), followed by EQP raw-raw-scale SM=0.3 (6.093).

Table 5.25.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.784	-0.026	6.305	6.302
IRT raw-scale	0.770	0.008	6.347	6.344
EQP raw-scale SM=0.0	0.769	0.078	6.077	6.075
EQP raw-scale SM=0.3	0.767	0.072	6.179	6.176
EQP raw-scale SM=1.0	0.768	0.065	6.157	6.154
EQP raw-raw-scale SM=0.0	0.769	0.036	6.101	6.098
EQP raw-raw-scale SM=0.3	0.769	0.030	6.096	6.093
EQP raw-raw-scale SM=1.0	0.769	0.052	6.103	6.101

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1014) over a large number of replications (i.e., 10,000).

Table 5.25.6: Comparison of Resampling Results 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 Neuro-QOL Pos Affect 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-scale SM=0.0 produced the smallest standard error, 0.678.

That is, the difference between the mean Neuro-QOL Pos Affect T-score and the mean equated TB Meaning T-score based on a similar sample of 75 cases is expected to be around ±1.36 (i.e., 2 × 0.678).

Table 5.25.6: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	-0.053	1.247	-0.006	0.869	-0.038	0.688
IRT raw-scale	0.012	1.251	0.035	0.884	0.015	0.700
EQP raw-scale SM=0.0	0.064	1.201	0.082	0.840	0.079	0.678
EQP raw-scale SM=0.3	0.066	1.216	0.065	0.841	0.067	0.689
EQP raw-scale SM=1.0	0.050	1.233	0.071	0.851	0.059	0.680
EQP raw-raw-scale SM=0.0	0.036	1.211	0.043	0.843	0.039	0.682
EQP raw-raw-scale SM=0.3	0.024	1.200	0.029	0.839	0.024	0.680
EQP raw-raw-scale SM=1.0	0.039	1.207	0.043	0.837	0.041	0.684

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

5.26 Neuro-QOL Positive Affect & Well-being and NIH Toolbox Positive Affect

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Emotion, namely the Neuro-QOL Positive Affect & Well-being item bank (23 items) and NIH Toolbox Positive Affect (20 items). Neuro-QOL Positive Affect & Well-being was scaled such that higher scores represent higher levels of Emotion. 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. Our sample consisted of 1,016 participants (N = 1,014 for participants with complete responses).

5.26.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 115 for Neuro-QOL Pos Affect and 100 for TB Pos Affect. Figures 5.26.1 and 5.26.2 graphically display the raw summed score distributions of the two measures. Figure 5.26.3 shows the distribution for the combined. Figure 5.26.4 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 Neuro-QOL Pos Affect and TB Pos Affect was 0.9. The correlations between the combined score and the measures were 0.97 and 0.97 for Neuro-QOL Pos Affect and TB Pos Affect, respectively.

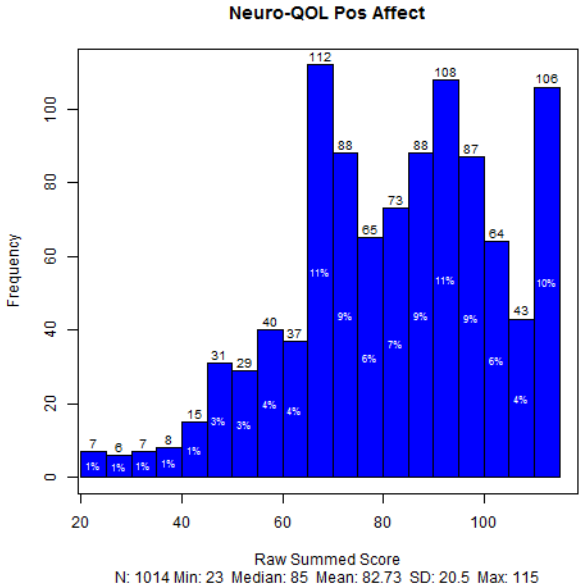


Figure 5.26.1: Raw Summed Score Distribution – Neuro-QOL Positive Affect & Well-being

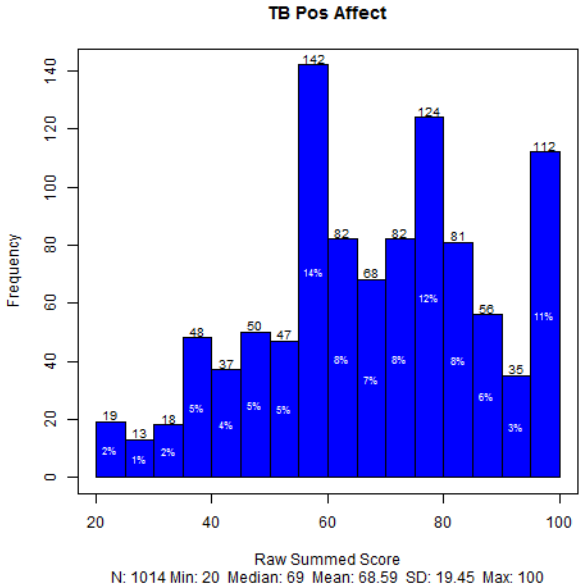


Figure 5.26.2: Raw Summed Score Distribution – NIH Toolbox Positive Affect

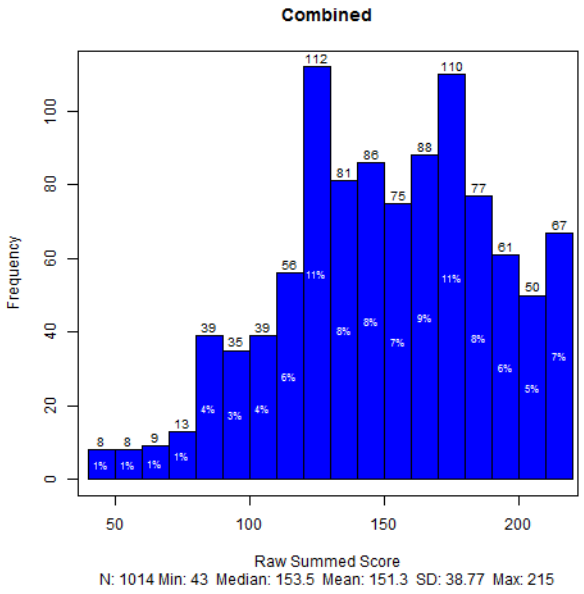


Figure 5.26.3: Raw Summed Score Distribution – Combined

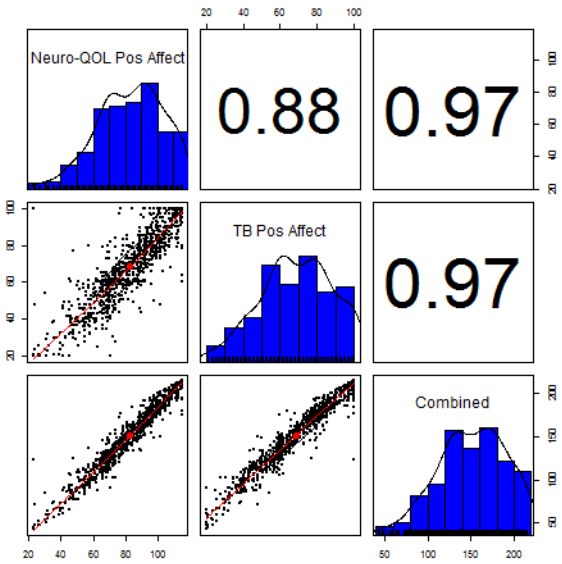


Figure 5.26.4: Scatter Plot Matrix of Raw Summed Scores

5.26.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.26.1 summarizes the results. For Neuro-QOL Pos Affect, Cron- bach's alpha internal consistency reliability estimate was 0.977 and adjusted (corrected for overlap) item-total correlations ranged from 0.733 to 0.859. For TB Pos Affect, alpha was 0.977 and adjusted item-total correlations ranged from 0.729 to 0.852. For the 43 items, alpha was 0.987 and adjusted item-total correlations ranged from 0.694 to 0.843.

Table 5.26.1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
Neuro-QOL Pos Affect	23	0.977	0.733	0.794	0.859
TB Pos Affect	20	0.977	0.729	0.817	0.852
Combined	43	0.987	0.694	0.789	0.843

5.26.3. 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 5.26.2 summarizes the model fit statistics. For Neuro-QOL Pos Affect, the fit statistics were as follows: CFI = 0.969, TLI = 0.966, and RMSEA = 0.104. For TB Pos Affect, CFI = 0.983, TLI = 0.981, and RMSEA = 0.092. For the 43

items, CFI = 0.954, TLI = 0.952, and RMSEA = 0.089.The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.26.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
Neuro-QOL Pos Affect	23	1016	0.969	0.966	0.104
TB Pos Affect	20	1016	0.983	0.981	0.092
Combined	43	1016	0.954	0.952	0.089

5.26.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 43 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 43 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 23 Neuro-QOL Pos Affect items served as anchor items to transform the item parameter estimates for the TB Pos Affect items onto the Neuro-QOL Pos Affect 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 5.26.3: IRT Linking Constants shows the additive (A) and multiplicative (B) transformation constants derived from the four linking methods. For fixed-parameter calibration, the item parameters for the Neuro-QOL Pos Affect items were constrained to their final bank values, while the TB Pos Affect items were calibrated, under the constraints imposed by the anchor items.

Table 5.26.3: IRT Linking Constants

	A	B
Mean/Mean	0.735	-0.024
Mean/Sigma	0.925	0.059
Haebara	0.904	0.050
Stocking-Lord	0.898	0.047

The item parameter estimates for the TB Pos Affect items were linked to the Neuro-QOL Pos Affect metric using the transformation constants shown in Table 5.26.3. The TB Pos Affect item parameter estimates from the fixed-parameter calibration are considered already on the Neuro-QOL Pos Affect metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for TB Pos Affect as shown in Figure 5.26.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.26.6 displays the differences on the vertical axis.

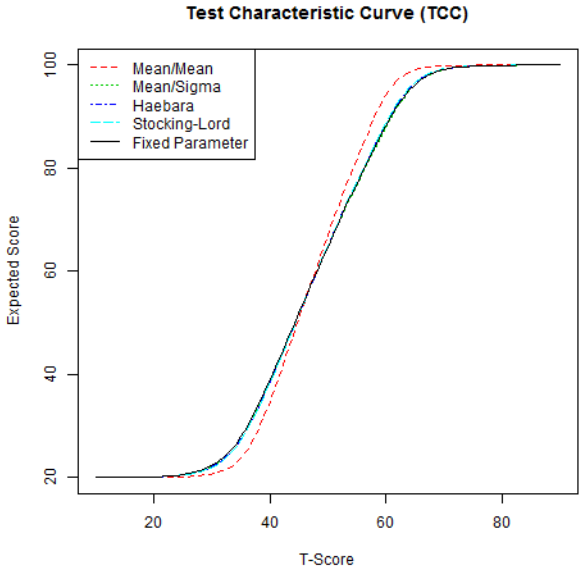


Figure 5.26.5: Test Characteristic Curves (TCC) from Different Linking Methods

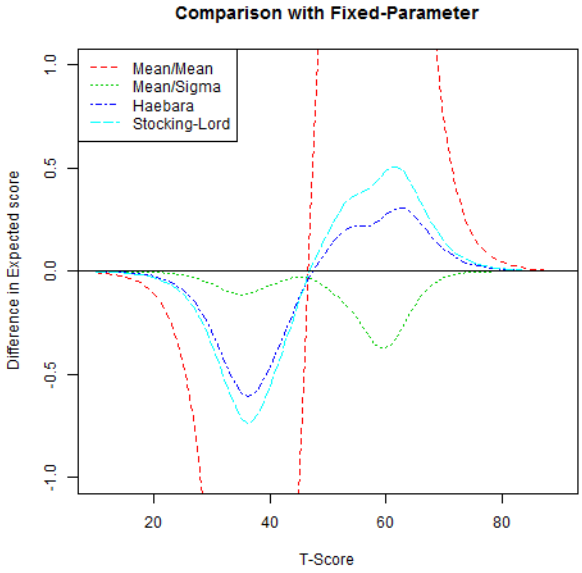


Figure 5.26.6: Difference in Test Characteristic Curves (TCC)

Table 5.26.4: Fixed-Parameter Estimates for shows the fixed-parameter calibration item parameter estimates for TB Pos Affect. The marginal reliability estimate for TB Pos Affect based on the item parameter estimates was 0.971. The marginal reliability estimates for Neuro-QOL Pos Affect and the combined set were 0.977 and 0.986, respectively. The slope parameter estimates for TB Pos Affect ranged from 2.29 to 3.9 with a mean of 3.35. The slope parameter estimates for Neuro-QOL Pos Affect ranged from 2.66 to 6.61 with a mean of 4.02. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.26.7 displays the scale information functions for Neuro-QOL Pos Affect, TB Pos Affect, and the combined set of 43. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.26.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.26.4: Fixed-Parameter Estimates for NIH Toolbox Positive Affect

a	cb1	cb2	cb3	cb4	NCAT
3.511	-1.288	-0.545	0.173	1.102	5
2.964	-1.662	-0.765	0.094	1.021	5
3.599	-1.211	-0.441	0.366	1.158	5
3.904	-1.398	-0.631	0.099	0.912	5
3.510	-1.313	-0.519	0.260	1.008	5
3.472	-1.195	-0.413	0.338	1.090	5
3.418	-1.176	-0.492	0.286	1.072	5
2.932	-1.510	-0.714	0.127	0.955	5
3.407	-1.435	-0.703	0.066	0.951	5
2.294	-1.581	-0.667	0.282	1.195	5
3.329	-1.332	-0.684	0.078	0.879	5
3.306	-1.174	-0.534	0.183	0.975	5
3.743	-1.276	-0.571	0.216	1.019	5

3.787	-1.129	-0.416	0.250	0.990	5
3.162	-1.124	-0.482	0.328	1.122	5
3.507	-1.401	-0.714	0.053	0.883	5
3.342	-1.354	-0.669	0.091	0.887	5
3.472	-1.271	-0.557	0.145	0.906	5
2.823	-1.222	-0.529	0.321	1.209	5
3.601	-1.122	-0.474	0.245	1.010	5

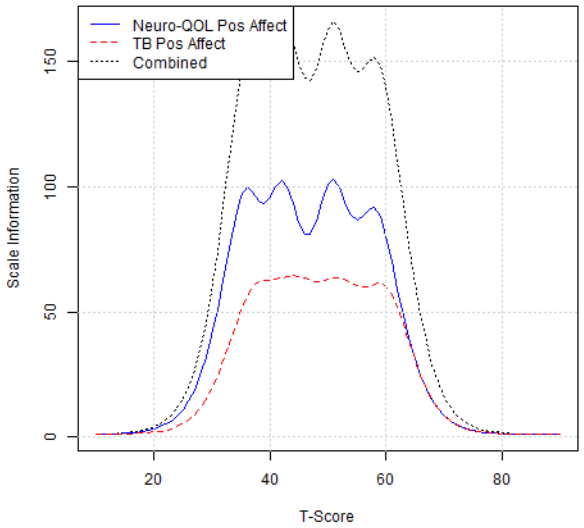


Figure 5.26.7: Comparison of Scale Information Functions

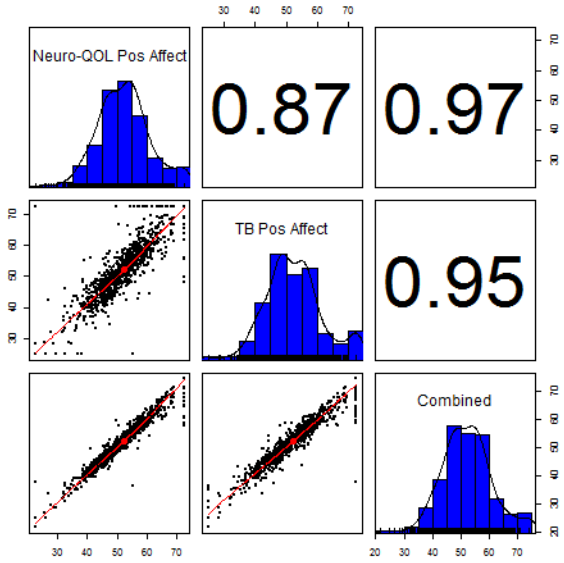


Figure 5.26.8: Comparison of IRT Scaled Scores

5.26.5. 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 TB Pos Affect to a scaled score on Neuro-QOL Pos Affect can be useful. Based on the TB Pos Affect item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 72 can be used to map simple raw summed scores from TB Pos Affect to T-score values linked to the Neuro-QOL Pos Affect metric. Each raw summed score point and corresponding Neuro-QOL 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.

5.26.6. Equipercentile Linking

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

percentile ranks as scores on TB Pos Affect. 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 5.26.9 displays the cumulative distribution functions of the measures. Figure 5.26.10 shows the equipercentile linking functions based on raw summed scores, from TB Pos Affect to Neuro-QOL Pos Affect. 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 73 and 74 show the equipercentile crosswalk tables. The result shown in Appendix Table 73 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 74 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).

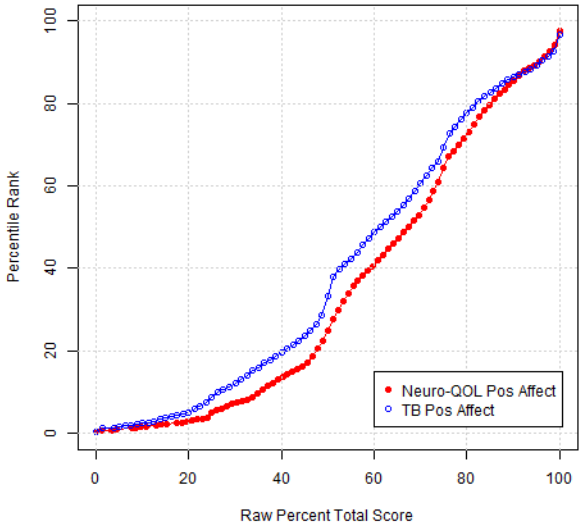


Figure 5.26.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

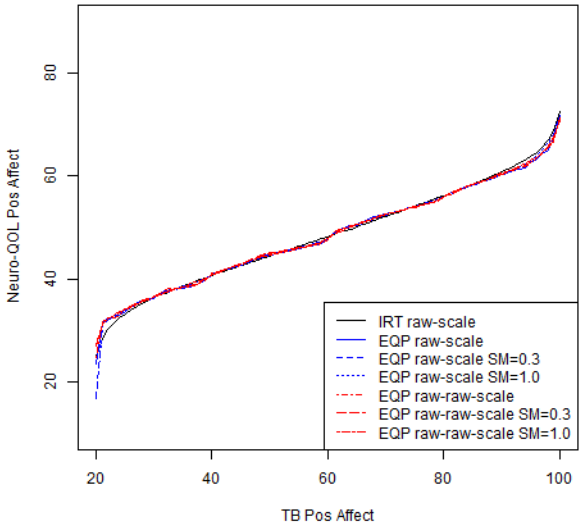


Figure 5.26.10: Equipercentile Linking Functions

5.26.7. 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 5.26.10, 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 5.26.5 reports four statistics summarizing the current sample in terms of the differences between the Neuro-QOL Pos Affect T-scores and TB Pos Affect scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.26.10), 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=1.0 produced the best result (0.87), followed by EQP raw-raw-scale SM=0.3 (0.869). 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 (4.556), followed by EQP raw-raw-scale SM=0.3 (4.568).

Table 5.26.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.869	-0.147	4.722	4.722
IRT raw-scale	0.868	-0.136	4.722	4.722
EQP raw-scale SM=0.0	0.866	0.026	4.681	4.678
EQP raw-scale SM=0.3	0.861	0.096	4.836	4.835
EQP raw-scale SM=1.0	0.861	0.102	4.828	4.826
EQP raw-raw-scale SM=0.0	0.867	0.045	4.635	4.633
EQP raw-raw-scale SM=0.3	0.869	0.049	4.570	4.568
EQP raw-raw-scale SM=1.0	0.870	0.078	4.558	4.556

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1014) over a large number of replications (i.e., 10,000).

Table 5.26.6: Comparison of Resampling Results 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 Neuro-QOL Pos Affect 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,

0.509. That is, the difference between the mean Neuro-QOL Pos Affect T-score and the mean equated TB Pos Affect T-score based on a similar sample of 75 cases is expected to be around ±1.02 (i.e., 2 × 0.509).

Table 5.26.6: Comparison of Resampling Results

Methods	Mean 25	SD 25	Mean 50	SD 50	Mean 75	SD 75
IRT pattern scoring	-0.141	0.930	-0.159	0.644	-0.147	0.524
IRT raw-scale	-0.125	0.937	-0.144	0.642	-0.138	0.528
EQP raw-scale SM=0.0	0.018	0.936	0.032	0.646	0.028	0.514
EQP raw-scale SM=0.3	0.113	0.953	0.087	0.662	0.096	0.531
EQP raw-scale SM=1.0	0.105	0.945	0.100	0.660	0.109	0.530
EQP raw-raw-scale SM=0.0	0.019	0.906	0.040	0.648	0.043	0.516
EQP raw-raw-scale SM=0.3	0.029	0.900	0.058	0.636	0.046	0.509
EQP raw-raw-scale SM=1.0	0.082	0.903	0.069	0.630	0.071	0.511

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

5.27. PROMIS Cognitive Function v2.0 and Neuro-QoL Cognitive Function v2.0

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Cognition, namely the PROMIS Cognitive Function item bank (32 items) and Neuro-QoL Cognitive Function (28 items). PROMIS Cognitive Function was scaled such that higher scores represent higher levels of Cognition. 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. Our sample consisted of 1,009 participants (N = 1,008 for participants with complete responses).

5.27.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 160 for PROMIS Cog Function and 140 for Neuro-QoL Cognitive Function. Figure 5.27.1 and Figure 5.27.2 graphically display the raw summed score distributions of the two measures. Figure 5.27.3 shows the distribution for the combined. Figure 5.27.4 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 Cog Function and Neuro-QoL Cognitive Function was 0.95. The disattenuated (corrected for unreliabilities) correlation between PROMIS Cog Function and Neuro-QoL Cognitive Function was 0.97. The correlations between the combined score and the measures were 0.99 and 0.98 for PROMIS Cog Function and Neuro-QoL Cognitive Function, respectively.

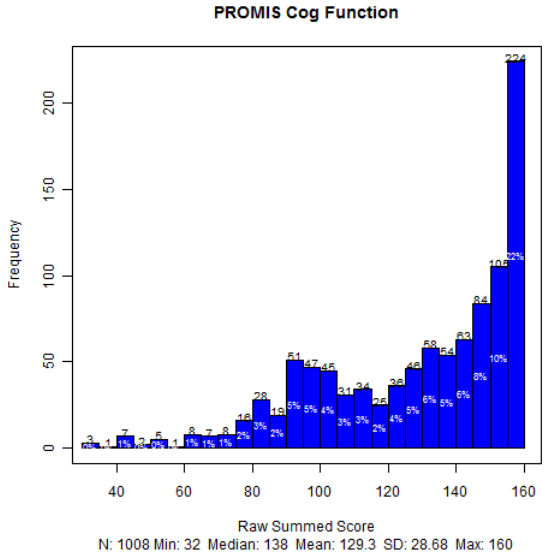


Figure 5.27.1: Raw Summed Score Distribution - PROMIS Cognitive Function

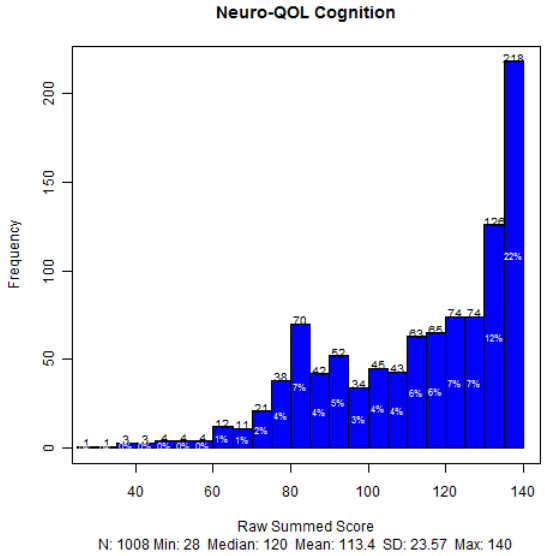


Figure 5.27.2: Raw Summed Score Distribution – Neuro-QoL Cognitive Function

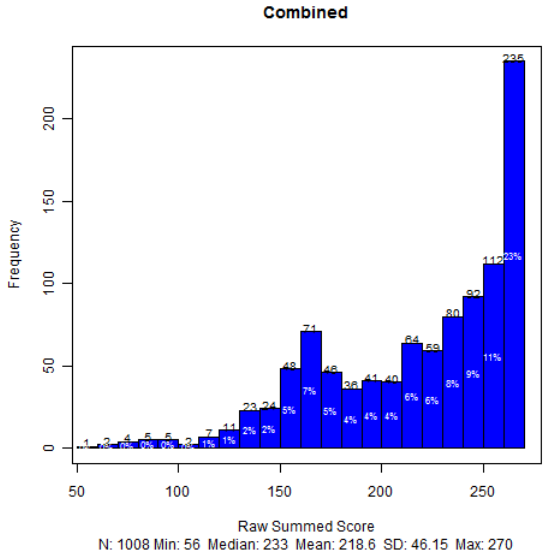


Figure 5.3.3: Raw Summed Score Distribution – Combined

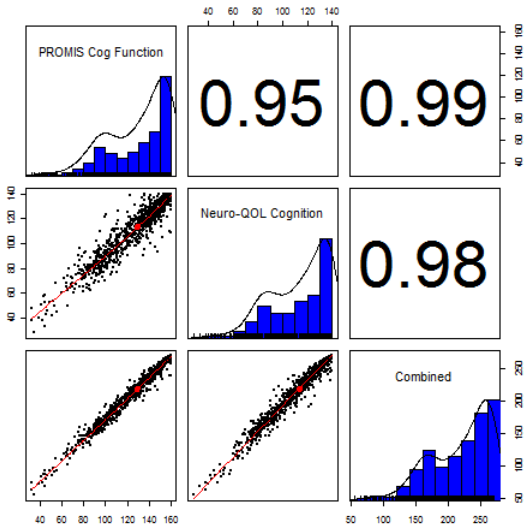


Figure 5.3.4: Scatter Plot Matrix of Raw Summed Scores

5.27.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 5.27.1 summarizes the results. For PROMIS Cog Function, Cronbach's alpha internal consistency reliability estimate was 0.982 and adjusted (corrected for overlap) item-total correlations ranged from 0.573 to 0.843. For Neuro-QoL Cognitive Function, alpha was 0.976 and adjusted item-total correlations ranged from 0.709 to 0.865. For the 54 items, alpha was 0.988 and adjusted item-total correlations ranged from 0.558 to 0.853.

Table 5.27.1: Classical Item Analysis

	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Adjusted (corrected for overlap) Item-total Correlation		
			Minimum	Mean	Maximum
PROMIS Cog Function	32	0.982	0.643	0.791	0.859
Neuro-QoL Cog Function	28	0.976	0.573	0.762	0.843
Combined	54	0.988	0.558	0.770	0.853

5.27.3. 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 5.27.2 summarizes the model fit statistics. For PROMIS Cog Function, the fit statistics were as follows: CFI = 0.976, TLI = 0.974, and RMSEA = 0.079. For Neuro-QOL Cognitive Function, CFI = 0.965, TLI = 0.962, and RMSEA =

0.097. For the 54 items, CFI = 0.962, TLI = 0.961, and RMSEA = 0.071. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 5.27.2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Cog Function	32	1009	0.976	0.974	0.079
Neuro-QoL Cog Function	28	1009	0.965	0.962	0.097
Combined	54	1009	0.962	0.961	0.071

5.27.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 54 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 54 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 32 PROMIS Cognitive Function items served as anchor items to transform the item parameter estimates for the Neuro-QoL Cognitive Function items onto the PROMIS Cognitive 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 5.27.3 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 Cognitive Function items were constrained to their final bank values, while the Neuro-QoL Cognitive Function items were calibrated, under the constraints imposed by the anchor items.

Table 5.27.3: IRT Linking Constants

	A	B
Mean/Mean	1.380	-0.493
Mean/Sigma	1.199	-0.567
Haebara	1.208	-0.561
Stocking-Lord	1.240	-0.548

The item parameter estimates for the Neuro-QoL Cognitive Function items were linked to the PROMIS Cognitive Function metric using the transformation constants shown in Table 5.27.3. The Neuro-QoL Cognitive Function item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Cognitive Function metric. Neuro-QoL Cognitive Function as shown in Figure 5.27.5. Using the fixed-parameter calibration as a basis we then examined the difference with each of the TCCs from the four linking methods. Figure 5.27.6 displays the differences on the vertical axis.

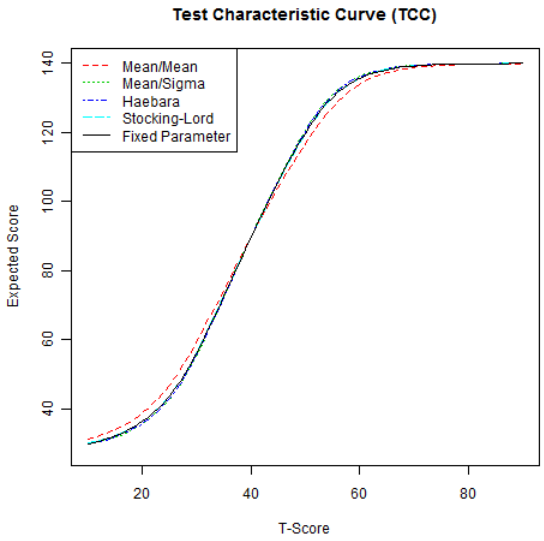


Figure 5.27.5: Test Characteristic Curves (TCC) from Different Linking Methods

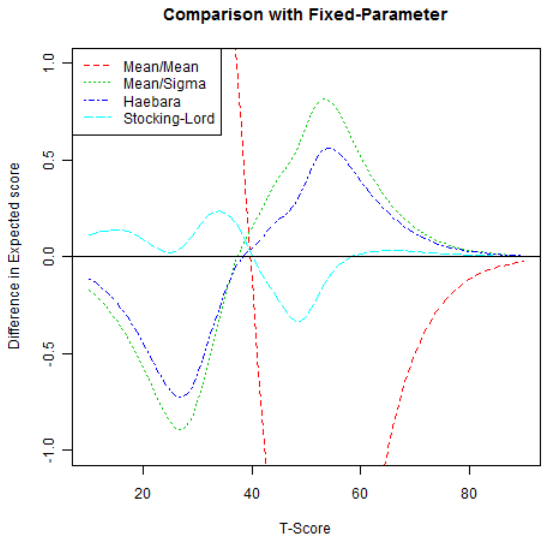


Figure 5.27.6: Difference in Test Characteristic Curves (TCC)

Table 5.27.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QoL Cognitive Function. The marginal reliability estimate for Neuro-QoL Cognitive Function based on the item parameter estimates was 0.951. The marginal reliability estimates for PROMIS Cognitive Function and the combined set were 0.952 and 0.97, respectively. The slope parameter estimates for Neuro-QoL Cognitive Function ranged from 1.48 to 3.93 with a mean of 2.56. The slope parameter estimates for PROMIS Cognitive Function ranged from 1.36 to 3.82 with a mean of 2.66. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.27.7 displays the scale information functions for PROMIS Cognitive Function, Neuro-QoL Cognitive Function, and the combined set of 54. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.27.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.27.4: Fixed-Parameter Calibration Item Parameter Estimates for Neuro-QoL Cognitive Function

a	cb1	cb2	cb3	cb4	NCAT
3.100	-1.890	-1.410	-0.810	-0.106	5
3.230	-1.860	-1.370	-0.753	-0.061	5
2.350	-2.250	-1.530	-0.729	0.105	5
2.590	-2.090	-1.430	-0.730	0.104	5
1.670	-2.610	-1.620	-0.534	0.795	5
3.020	-1.960	-1.360	-0.701	0.007	5
2.515	-2.279	-1.742	-0.870	0.292	5
1.898	-2.381	-1.558	-0.452	0.856	5
2.284	-2.358	-1.605	-0.558	0.529	5
3.253	-2.139	-1.385	-0.626	0.341	5
3.158	-1.981	-1.458	-0.617	0.183	5
2.527	-2.448	-1.640	-0.851	0.006	5
3.736	-1.948	-1.410	-0.763	0.053	5
3.932	-1.953	-1.410	-0.794	-0.071	5
3.322	-1.898	-1.414	-0.608	0.204	5
3.471	-1.929	-1.401	-0.746	-0.035	5
3.184	-1.962	-1.361	-0.695	0.030	5
3.727	-2.036	-1.453	-0.826	-0.117	5
1.477	-3.504	-2.269	-1.311	-0.371	5
1.768	-3.216	-1.871	-1.094	-0.197	5
1.994	-2.780	-1.935	-1.052	-0.180	5
2.004	-3.020	-1.876	-0.969	-0.129	5
1.908	-2.939	-1.862	-0.853	0.223	5
2.021	-3.063	-1.830	-0.910	-0.181	5
1.789	-2.865	-1.676	-0.813	0.299	5
1.866	-3.036	-1.818	-0.833	0.446	5
1.754	-2.567	-1.505	-0.608	0.680	5
2.270	-2.748	-1.802	-0.896	0.095	5

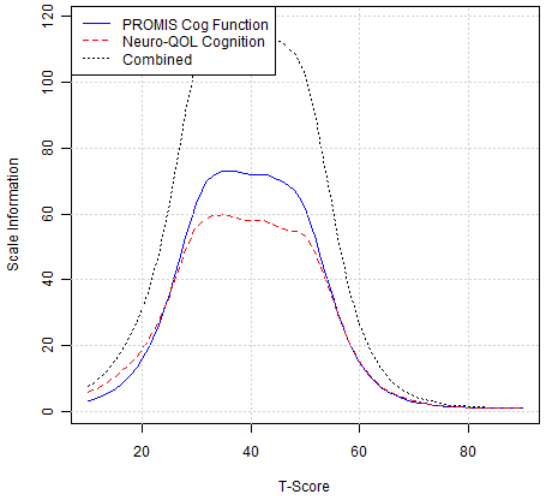


Figure 5.27.7: Comparison of Scale Information Functions

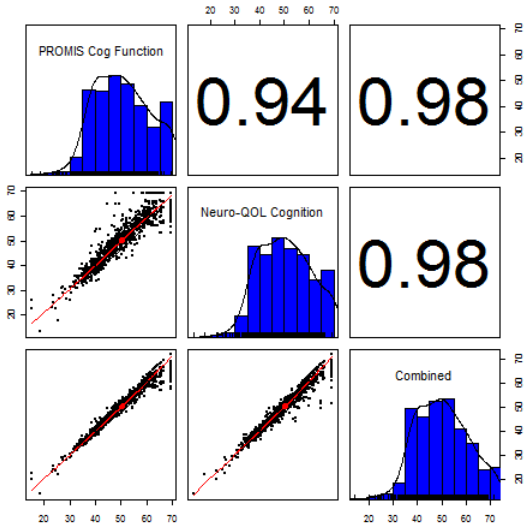


Figure 5.27.8: Comparison of IRT Scaled Scores

5.27.5. 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 Cognitive Function to a scaled score on PROMIS Cognitive Function can be useful. Based on the Neuro-QoL Cognitive Function item parameters derived from the fixed- parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 75 can be used to map simple raw summed scores Neuro-QoL Cognitive Function to T-score values linked to the PROMIS Cognitive Function metric. (This is now equivalent to the NeuroQoL Cognitive Function T-score 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.

5.27.6. Equipercentile Linking

We mapped each raw summed score point on Neuro-QoL Cognitive Function to a corresponding scaled score on PROMIS Cognitive Function by identifying scores on PROMIS Cognitive Function that have the same percentile ranks as scores on Neuro-QoL Cognitive Function. 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 5.27.9 displays the cumulative distribution functions of the measures. Figure 5.27.10 shows the

equipercentile linking functions based on raw summed scores, from Neuro-QoL Cognitive Function to PROMIS Cognitive 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 76 and Appendix Table 77 show the equipercentile crosswalk tables. The result shown in Appendix Table 76 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 77 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).

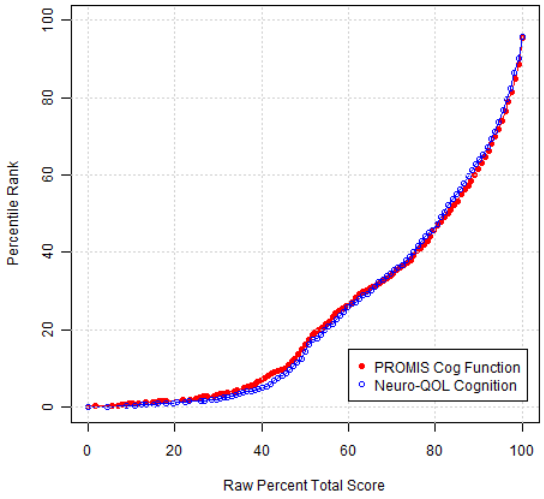


Figure 5.27.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

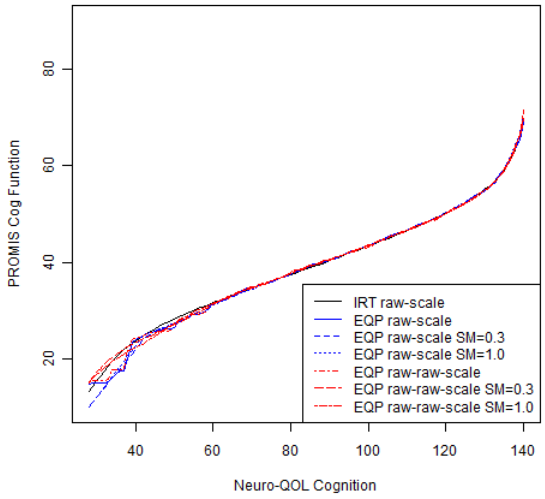


Figure 5.27.10: Equipercentile Linking Functions

5.27.7. 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 5.27.10, 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 5.27.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Cognitive Function T-scores and Neuro-QoL Cognitive Function scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.27.10), 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 (0.942), followed by IRT raw-scale (0.934). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (3.634), followed by IRT raw-scale (3.868).

Table 5.27.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.942	0.062	3.635	3.634
IRT raw-scale	0.934	-0.002	3.877	3.875
EQP raw-scale SM=0.0	0.934	-0.023	3.870	3.868
EQP raw-scale SM=0.3	0.933	-0.160	3.960	3.962
EQP raw-scale SM=1.0	0.934	-0.157	3.893	3.894
EQP raw-raw-scale SM=0.0	0.934	-0.131	3.894	3.894
EQP raw-raw-scale SM=0.3	0.934	-0.118	3.889	3.888
EQP raw-raw-scale SM=1.0	0.931	-0.177	4.029	4.031

To examine the bias and standard error of the linking results, a resampling study was used. In this procedure, small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1008) 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 Cognitive 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, 0.405. That is, the difference between the mean PROMIS Cognitive Function T-score and the mean equated Neuro-QoL Cognitive Function T-score based on a similar sample of 75 cases is expected to be around ±0.81 (i.e., 2 × 0.405).

Table 5.27.6: 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	0.074	0.712	0.063	0.506	0.071	0.405
IRT raw-scale	-0.007	0.763	0.003	0.536	0.005	0.432
EQP raw-scale SM=0.0	-0.026	0.763	-0.021	0.527	-0.028	0.431
EQP raw-scale SM=0.3	-0.167	0.790	-0.170	0.542	-0.156	0.443
EQP raw-scale SM=1.0	-0.165	0.774	-0.153	0.529	-0.154	0.434
EQP raw-raw-scale SM=0.0	-0.143	0.767	-0.130	0.525	-0.129	0.431
EQP raw-raw-scale SM=0.3	-0.114	0.759	-0.112	0.533	-0.116	0.432
EQP raw-raw-scale SM=1.0	-0.179	0.800	-0.185	0.555	-0.177	0.450

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

6.0 References

Beck, A.T., & Steer, R.A. (1993). *Manual for the Beck Depression Inventory*. San Antonio: Psychological Corporation.

Beck, A.T., Steer, R.A., Ball, R., & Ranieri, W.F. (1996). Comparison of Beck Depression Inventories-Ia and-Ii in Psychiatric Outpatients. *Journal of Personality Assessment, 67*(3), 588-597. doi: 10.1207/s15327752jpa6703_13

Beck, A.T., Steer, R.A., & Brown, G.K. (1996). *Manual for the Beck Depression Inventory—Ii*. San Antonio, TX: Psychological Corporation.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry, 4*, 561-571.

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).

Buysse, D.J., Reynolds, C.F., III, Monk, T.H., Berman, S.R., & Kupfer, D.J. (1989). The Pittsburgh Sleep Quality Index: A New Instrument for Psychiatric Practice and Research. *Psychiatry Research, 28*(2), 193-213.

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.

Choi, S.W., Podrabsky, T., McKinney, N., Schalet, B.D., Cook, K.F., & Cella, D. (2012). *Prosetta Stone (Tm) Analysis Report: A Rosetta Stone for Patient Reported Outcomes* (Vol. 1). Chicago, IL: Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University.

Dorans, N.J. (2007). Linking Scores from Multiple Health Outcome Instruments. *Quality of Life Research, 16*(Supplement 1), 85-94. doi: 10.2307/40212575

Gershon, R.C. (2007). NIH Toolbox: Assessment of Neurological and Behavioral Function. NIH (Contract Hhs-N-260-2006 00007-C). Retrieved November 1, 2013, from http://www.nihtoolbox.org

Kazis, L.E., Miller, D.R., Clark, J.A., Skinner, K.M., Lee, A., Ren, X.S., . . . Ware, J.E.J. (2004). Improving the Response Choices on the Veterans SF-36 Health Survey Role Functioning Scales: Results from the Veterans Health Study. *Journal of Ambulatory Care Management, 27*(3), 263-280.

Kessler, R.C., Green, J., Adler, L.A., & et al. (2010). Structure and Diagnosis of Adult Attention-Deficit/Hyperactivity Disorder: Analysis of Expanded Symptom Criteria from the Adult Adhd Clinical Diagnostic Scale. *Archives of General Psychiatry, 67*(11), 1168-1178. doi: 10.1001/archgenpsychiatry.2010.146

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. (2003). The Patient Health Questionnaire-2: Validity of a Two-Item Depression Screener. *Medical Care, 41*(11), 1284-1292.

Lai, J.-S., Butt, Z., Zelko, F., Cella, D., Krull, K., Kieran, M., & Goldman, S. (2011). Development of a Parent-Report Cognitive Function Item Bank Using Item Response Theory and Exploration of Its Clinical Utility in Computerized Adaptive Testing. *Journal of Pediatric Psychology, 36*(7), 766-779.

Lai, J.-S., Zelko, F., Krull, K., Cella, D., Nowinski, C., Manley, P., & Goldman, S. (2013). Parent-Reported Cognition of Children with Cancer and Its Potential Clinical Usefulness. *Quality of Life Research, Epub ahead of print*, 1-10. doi: 10.1007/s11136-013-0548-9

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

Neuro-Qol - Quality of Life Outcomes in Neurological Disorders. (2008). Retrieved May 1, 2010, from http://www.neuroqol.org

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. Psychometrika Monograph Supplement, No. 17* Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Selim, A., Rogers, W., Fleishman, J., Qian, S., Fincke, B., Rothendler, J., & Kazis, L. (2009). Updated U.S. Population Standard for the Veterans Rand 12-Item Health Survey (VR-12). *Quality of Life Research, 18*(1), 43-52. doi: 10.1007/s11136-008-9418-2

Ware, J.E., Kosinski, M., & Dewey, J.E. (2000). *How to Score Version 2 of the SF-36 Health Survey*. Lincoln, R.I.: QualityMetric.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The Panas Scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Zigmond, A.S., & Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67*(6), 361-370.

7.0 Appendix

Appendix Table 1: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for FACT-Cog Abilities to PROMIS Cognitive Function - Abilities (PROsetta Stone Wave 2 Study) – RECOMMENDED

FACT-Cog Abilities Score	PROMIS T-score	SE
0	22.5	4.0
1	25.7	3.2
2	27.6	2.9
3	29.1	2.6
4	30.4	2.4
5	31.5	2.3
6	32.5	2.2
7	33.5	2.2
8	34.4	2.1
9	35.2	2.1
10	36.1	2.1
11	36.9	2.1
12	37.7	2.1
13	38.5	2.1
14	39.3	2.1
15	40.1	2.1
16	41.0	2.1
17	41.8	2.1
18	42.6	2.1
19	43.5	2.1
20	44.3	2.2
21	45.2	2.2
22	46.1	2.2
23	47.0	2.2
24	47.9	2.2
25	48.8	2.2
26	49.8	2.2
27	50.7	2.2
28	51.7	2.2
29	52.8	2.2
30	53.9	2.3
31	55.1	2.4
32	56.5	2.5
33	58.0	2.8
34	59.9	3.1
35	62.4	3.7
36	67.0	5.1

Appendix Table 2: Direct (Raw to Scale) Equipercentile Crosswalk Table – From FACT-Cog Abilities to PROMIS Cognitive Function – Abilities – Note: Table 1 is recommended.

FACT-Cog Abilities Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	27	28	28	2.09
1	31	30	30	2.35
2	32	32	32	1.04
3	34	33	33	0.53
4	34	34	34	0.55
5	34	34	34	0.43
6	35	35	35	0.44
7	35	35	35	0.43
8	36	36	36	0.29
9	36	36	36	0.30
10	37	37	37	0.43
11	38	38	38	0.39
12	38	38	38	0.38
13	39	39	39	0.30
14	40	40	40	0.26
15	40	40	41	0.27
16	41	41	41	0.22
17	42	42	42	0.17
18	43	43	43	0.35
19	44	44	44	0.33
20	45	45	45	0.54
21	46	46	46	0.27
22	46	46	46	0.25
23	47	47	47	0.35
24	48	48	48	0.26
25	49	49	49	0.29
26	50	50	50	0.20
27	51	51	51	0.34
28	52	52	52	0.41
29	54	54	54	0.34
30	55	55	55	0.53
31	56	56	56	0.25
32	57	57	57	0.35
33	59	59	58	0.25
34	60	59	60	0.36
35	61	61	62	0.36
36	69	69	68	0.12

Appendix Table 3: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From FACT-Cog Abilities to PROMIS Cognitive Function – Abilities –Note: Table 1 is recommended.

FACT-Cog Abilities Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	27	28	29
1	31	31	30
2	32	32	32
3	34	33	33
4	34	34	33
5	34	34	34
6	35	35	35
7	35	35	35
8	36	36	36
9	36	36	36
10	37	37	37
11	38	38	38
12	38	38	38
13	39	39	39
14	40	40	40
15	40	40	40
16	41	41	41
17	42	42	42
18	43	43	43
19	44	44	44
20	45	45	45
21	46	46	46
22	46	46	46
23	47	47	47
24	48	48	48
25	49	49	49
26	50	50	50
27	51	51	51
28	52	52	52
29	54	54	53
30	55	55	55
31	56	56	56
32	58	57	57
33	59	58	58
34	60	60	60
35	61	62	62
36	67	67	67

Appendix Table 4: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for FACT-Cog Perceived Cognitive Impairment to PROMIS Cognitive Function (PROsetta Stone Wave 2 Study) - RECOMMENDED

FACT-Cog Perceived Cognitive Impairment Score	PROMIS T-score	SE	FACT-Cog Perceived Cognitive Impairment Score	PROMIS T-score	SE
80	14.7	2.7	39	38.5	1.7
79	16.3	2.8	38	38.9	1.7
78	17.6	2.7	37	39.4	1.7
77	18.8	2.6	36	39.8	1.8
76	19.9	2.5	35	40.2	1.8
75	20.8	2.4	34	40.6	1.8
74	21.7	2.3	33	41.1	1.8
73	22.5	2.2	32	41.5	1.8
72	23.2	2.1	31	42.0	1.8
71	23.9	2.0	30	42.4	1.8
70	24.5	2.0	29	42.9	1.8
69	25.1	1.9	28	43.3	1.8
68	25.7	1.9	27	43.8	1.8
67	26.2	1.9	26	44.2	1.8
66	26.7	1.9	25	44.7	1.8
65	27.3	1.8	24	45.2	1.8
64	27.8	1.8	23	45.7	1.8
63	28.2	1.8	22	46.2	1.8
62	28.7	1.8	21	46.7	1.8
61	29.2	1.8	20	47.2	1.9
60	29.6	1.8	19	47.7	1.9
59	30.1	1.8	18	48.3	1.9
58	30.5	1.8	17	48.8	1.9
57	31.0	1.8	16	49.4	1.9
56	31.4	1.8	15	50.0	1.9
55	31.9	1.8	14	50.6	1.9
54	32.3	1.7	13	51.2	1.9
53	32.7	1.7	12	51.8	2.0
52	33.1	1.7	11	52.5	2.0
51	33.5	1.7	10	53.2	2.1
50	34.0	1.7	9	53.9	2.1
49	34.4	1.7	8	54.7	2.2
48	34.8	1.7	7	55.6	2.3
47	35.2	1.7	6	56.5	2.4
46	35.6	1.7	5	57.6	2.6
45	36.0	1.7	4	58.8	2.8
44	36.4	1.7	3	60.3	3.1
43	36.9	1.7	2	62.1	3.5
42	37.3	1.7	1	64.6	4.0
41	37.7	1.7	0	68.6	5.2
40	38.1	1.7			

Appendix Table 5: Direct (Raw to Scale) Equipercentile Crosswalk Table – From FACT-Cog Perceived Cognitive Impairment to PROMIS Cognitive Function – Abilities - Note: Table 4 is recommended.

FACT-Cog Perceived Cognitive Impairment Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
80	13	10	10	2.00
79	14	12	12	2.00
78	19	14	14	2.00
77	20	16	16	1.22
76	21	17	18	1.22
75	21	19	20	1.27
74	22	20	21	0.71
73	22	21	22	0.72
72	22	22	23	0.66
71	22	22	23	0.66
70	22	23	24	2.45
69	23	24	24	2.45
68	24	24	25	2.83
67	24	25	25	2.83
66	26	26	26	1.05
65	26	27	27	1.05
64	27	27	27	1.73
63	29	28	28	1.73
62	30	29	28	0.53
61	30	29	29	0.51
60	30	30	29	0.51
59	30	30	30	0.58
58	30	31	30	2.26
57	32	31	31	0.58
56	32	31	31	0.57
55	32	32	32	0.61
54	32	32	32	0.61
53	33	33	32	0.50
52	33	33	33	0.42
51	33	33	33	0.40
50	34	34	34	0.53
49	34	34	34	0.58
48	35	35	34	0.42
47	35	35	35	0.42
46	35	35	35	0.42
45	36	36	36	0.32
44	36	36	36	0.31
43	36	37	37	0.32
42	37	37	37	0.43
41	38	38	37	0.20

40	38	38	38	0.21
39	38	38	38	0.27
38	39	39	39	0.25
37	39	39	39	0.24
36	40	40	40	0.34
35	40	40	40	0.32
34	40	40	40	0.31
33	41	41	41	0.36
32	41	41	41	0.34
31	42	42	42	0.41
30	42	42	42	0.40
29	43	43	43	0.42
28	43	43	43	0.39
27	43	44	44	0.39
26	44	44	44	0.58
25	45	45	45	0.59
24	45	45	45	0.57
23	46	46	46	0.35
22	46	46	46	0.33
21	47	47	47	0.47
20	48	48	47	0.26
19	48	48	48	0.26
18	49	49	49	0.40
17	49	49	49	0.40
16	50	50	50	0.22
15	50	50	50	0.21
14	51	51	51	0.59
13	52	51	51	0.25
12	52	52	52	0.24
11	52	52	52	0.23
10	53	53	53	0.53
9	54	54	54	0.37
8	54	54	55	0.36
7	55	55	55	0.34
6	56	56	57	0.31
5	58	58	58	0.44
4	60	59	59	0.46
3	61	61	61	0.44
2	63	63	64	0.34
1	65	66	66	0.30
0	71	71	71	0.11

Appendix Table 6: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From FACT-Cog Perceived Cognitive Impairment to PROMIS Cognitive Function – Abilities - Note: Table 4 is recommended

FACT-Cog Perceived Cognitive Impairment Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
80	13	13	13
79	14	14	14
78	19	16	16
77	20	17	17
76	21	18	18
75	21	19	19
74	22	20	20
73	22	21	21
72	22	21	22
71	22	22	22
70	23	23	23
69	23	24	24
68	23	25	25
67	24	25	25
66	26	26	26
65	26	27	27
64	27	28	27
63	29	28	28
62	30	29	28
61	30	29	29
60	30	30	29
59	30	30	30
58	30	31	30
57	31	31	31
56	32	32	31
55	32	32	32
54	32	32	32
53	32	33	33
52	33	33	33
51	34	34	33
50	34	34	34
49	34	34	34
48	35	34	35
47	35	35	35
46	35	35	36
45	36	36	36

44	36	36	36
43	36	37	37
42	37	37	37
41	38	38	38
40	38	38	38
39	38	38	38
38	39	39	39
37	39	39	39
36	40	40	40
35	40	40	40
34	40	40	41
33	41	41	41
32	41	41	42
31	42	42	42
30	42	42	42
29	42	42	43
28	43	43	43
27	43	44	44
26	44	44	44
25	44	44	45
24	45	45	45
23	46	46	46
22	46	46	46
21	47	47	47
20	48	48	47
19	48	48	48
18	49	49	48
17	49	49	49
16	50	50	49
15	50	50	50
14	51	51	50
13	52	51	51
12	52	52	52
11	53	52	52
10	53	53	53
9	54	54	54
8	54	55	55
7	55	56	56
6	57	56	57
5	58	58	58
4	59	59	59
3	61	61	61
2	63	63	63

1	66	66	66
0	70	70	71

Appendix Table 7: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QOL Applied Cognition -General Concerns to PROMIS Cognitive Function v2.0 (PROsetta Stone Wave 2 Study) – RECOMMENDED

Note: In 2014, the two Neuro-QoL Applied Cognition banks -- General Concerns and Executive Function – were merged into a single bank called Neuro-QoL Cognitive Function. This new bank was linked via common items to the PROMIS Cognitive Function v2.0 bank, so that the T-scores from either instrument are on the same metric. See Report 5.27 (from vol. 2) for details on the link with Neuro-QoL Cognitive Function v2.0.

Neuro-QOL Appl Cog General Concerns T-Score	Neuro-QOL Appl Cog General Concerns Raw Score	PROMIS T-score	SE
16.3	18	16.7	3.2
18.3	19	18.8	3
19.5	20	20.4	2.8
20.7	21	21.6	2.6
21.6	22	22.7	2.4
22.5	23	23.7	2.3
23.2	24	24.5	2.2
23.8	25	25.3	2.1
24.4	26	26	2
25.0	27	26.7	1.9
25.5	28	27.3	1.9
26.0	29	27.9	1.8
26.4	30	28.4	1.8
26.9	31	29	1.8
27.3	32	29.5	1.7
27.7	33	30	1.7
28.1	34	30.5	1.7
28.5	35	30.9	1.7
28.9	36	31.4	1.7
29.2	37	31.9	1.7
29.6	38	32.3	1.7
30.0	39	32.8	1.7
30.3	40	33.3	1.7
30.7	41	33.7	1.7
31.1	42	34.2	1.7
31.4	43	34.6	1.7
31.8	44	35	1.7
32.2	45	35.5	1.7
32.5	46	35.9	1.7
32.9	47	36.4	1.7
33.3	48	36.8	1.7
33.6	49	37.3	1.7

34.0	50	37.7	1.7
34.4	51	38.2	1.7
34.7	52	38.7	1.7
35.1	53	39.1	1.7
35.5	54	39.6	1.7
35.8	55	40	1.7
36.2	56	40.5	1.7
36.6	57	41	1.7
37.0	58	41.5	1.7
37.3	59	41.9	1.7
37.7	60	42.4	1.7
38.1	61	42.9	1.7
38.5	62	43.4	1.7
38.9	63	43.9	1.7
39.3	64	44.4	1.7
39.7	65	44.9	1.7
40.0	66	45.4	1.7
40.4	67	45.9	1.7
40.8	68	46.4	1.7
41.2	69	46.9	1.7
41.7	70	47.4	1.7
42.1	71	47.9	1.7
42.5	72	48.5	1.7
42.9	73	49	1.7
43.4	74	49.6	1.8
43.8	75	50.2	1.8
44.3	76	50.7	1.8
44.8	77	51.3	1.8
45.3	78	52	1.8
45.8	79	52.6	1.9
46.4	80	53.3	1.9
47.0	81	54	2
47.6	82	54.8	2.1
48.4	83	55.7	2.2
49.2	84	56.6	2.3
50.2	85	57.7	2.5
51.3	86	59	2.8
52.7	87	60.5	3.1
54.5	88	62.4	3.5
57.0	89	64.9	4.1
62.5	90	68.9	5.2

Appendix Table 8: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QOL Applied Cognition -General Concerns to PROMIS Cognitive Function – Abilities - Note: Table 7 is recommended.

Neuro-QOL Appl Cog General Concerns T-Score	Neuro-QOL Appl Cog General Concerns Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
16.3	18	15	12	12	2.45
18.3	19	21	17	18	0.35
19.5	20	22	18	19	0.71
20.7	21	22	20	21	0.71
21.6	22	22	21	22	0.35
22.5	23	22	22	23	0.35
23.2	24	22	23	23	0.61
23.8	25	24	23	24	2.45
24.4	26	24	24	24	2
25.0	27	25	25	25	2
25.5	28	26	26	26	0.82
26.0	29	26	26	26	0.77
26.4	30	27	27	27	1
26.9	31	27	28	28	0.79
27.3	32	28	28	28	1
27.7	33	29	29	29	1
28.1	34	30	30	30	0.35
28.5	35	30	30	30	0.41
28.9	36	31	31	31	1.54
29.2	37	32	32	31	0.43
29.6	38	32	32	32	0.44
30.0	39	33	32	32	0.34
30.3	40	33	33	33	0.33
30.7	41	33	33	33	0.33
31.1	42	33	33	33	0.42
31.4	43	34	34	34	0.49
31.8	44	34	34	34	0.49
32.2	45	35	35	35	0.37
32.5	46	35	35	35	0.34
32.9	47	36	36	36	0.27
33.3	48	36	36	36	0.28
33.6	49	37	37	37	0.35
34.0	50	37	37	37	0.35

34.4	51	38	38	38	0.15
34.7	52	38	38	38	0.15
35.1	53	38	38	38	0.16
35.5	54	39	39	39	0.2
35.8	55	40	39	39	0.29
36.2	56	40	40	40	0.26
36.6	57	41	40	40	0.29
37.0	58	41	41	41	0.27
37.3	59	41	41	41	0.27
37.7	60	42	42	42	0.31
38.1	61	42	42	42	0.3
38.5	62	43	43	43	0.3
38.9	63	43	43	43	0.28
39.3	64	44	44	44	0.39
39.7	65	44	44	44	0.39
40.0	66	45	45	45	0.37
40.4	67	46	45	45	0.23
40.8	68	46	46	46	0.22
41.2	69	46	46	46	0.23
41.7	70	47	47	47	0.39
42.1	71	48	48	48	0.22
42.5	72	48	48	48	0.21
42.9	73	49	49	49	0.3
43.4	74	50	49	49	0.17
43.8	75	50	50	50	0.17
44.3	76	50	50	50	0.17
44.8	77	51	51	51	0.49
45.3	78	52	52	51	0.22
45.8	79	52	52	52	0.21
46.4	80	52	53	53	0.21
47.0	81	53	53	53	0.46
47.6	82	54	54	54	0.3
48.4	83	55	55	55	0.29
49.2	84	56	56	56	0.27
50.2	85	56	57	57	0.25
51.3	86	58	58	58	0.34
52.7	87	61	60	61	0.4
54.5	88	63	63	63	0.33
57.0	89	65	66	66	0.27
62.5	90	71	71	71	0.1

Appendix Table 9: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QOL Applied Cognition -General Concerns to PROMIS Cognitive Function – Abilities - Note: Table 7 is recommended.

Neuro-QOL Appl Cog General Concerns T-Score	Neuro-QOL Appl Cog General Concerns Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
16.3	18	15	15	14
18.3	19	21	19	18
19.5	20	22	20	19
20.7	21	22	21	20
21.6	22	22	22	22
22.5	23	22	22	23
23.2	24	23	23	24
23.8	25	23	24	24
24.4	26	24	24	25
25.0	27	25	25	26
25.5	28	26	25	26
26.0	29	26	26	27
26.4	30	27	27	28
26.9	31	27	28	28
27.3	32	27	28	29
27.7	33	29	29	29
28.1	34	30	30	30
28.5	35	30	30	30
28.9	36	31	31	31
29.2	37	32	32	31
29.6	38	32	32	32
30.0	39	32	32	32
30.3	40	33	33	33
30.7	41	33	33	33
31.1	42	34	34	34
31.4	43	34	34	34
31.8	44	34	34	34
32.2	45	35	35	35
32.5	46	35	35	35
32.9	47	36	36	36
33.3	48	36	36	36
33.6	49	37	37	37
34.0	50	37	37	37

34.4	51	38	38	38
34.7	52	38	38	38
35.1	53	38	38	38
35.5	54	39	39	39
35.8	55	40	39	39
36.2	56	40	40	40
36.6	57	41	40	40
37.0	58	41	41	41
37.3	59	41	41	41
37.7	60	42	42	42
38.1	61	42	42	42
38.5	62	43	43	43
38.9	63	43	43	43
39.3	64	44	44	44
39.7	65	44	44	44
40.0	66	45	45	45
40.4	67	45	45	45
40.8	68	46	46	46
41.2	69	46	46	46
41.7	70	47	47	47
42.1	71	47	47	47
42.5	72	48	48	48
42.9	73	49	49	48
43.4	74	49	49	49
43.8	75	50	50	50
44.3	76	50	50	50
44.8	77	51	51	51
45.3	78	52	52	52
45.8	79	52	52	52
46.4	80	53	53	53
47.0	81	54	53	54
47.6	82	54	54	54
48.4	83	55	55	55
49.2	84	56	56	56
50.2	85	57	57	57
51.3	86	58	58	59
52.7	87	60	60	60
54.5	88	63	63	62
57.0	89	65	66	65
62.5	90	70	70	70

Appendix Table 10: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Peds PCF Short Form and PROMIS Cognitive Function (PROsetta Stone Wave 2 Study) - RECOMMENDED

Peds PCF Short Form Raw Score	PROMIS T-score	SE
7	15.9	3.5
8	18.0	3.7
9	20.1	3.8
10	22.1	3.8
11	23.9	3.7
12	25.6	3.7
13	27.3	3.6
14	28.8	3.6
15	30.3	3.6
16	31.7	3.6
17	33.1	3.6
18	34.5	3.6
19	35.8	3.6
20	37.2	3.6
21	38.6	3.6
22	39.9	3.6
23	41.3	3.6
24	42.7	3.6
25	44.1	3.6
26	45.6	3.7
27	47.1	3.7
28	48.7	3.7
29	50.3	3.8
30	52.1	3.8
31	54.0	4.0
32	56.2	4.1
33	58.7	4.4
34	61.7	4.7
35	66.4	5.7

Appendix Table 11: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Peds PCF Short Form to PROMIS Cognitive Function – Note: Table 10 is recommended.

Peds PCF Short Form Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
7	14	11	11	2.45
8	20	14	14	1.22
9	20	17	17	1.22
10	21	20	20	1.22
11	22	21	22	0.67
12	24	24	25	4.47
13	26	26	27	1.05
14	30	29	29	0.53
15	32	31	31	0.47
16	33	33	32	0.44
17	34	34	34	0.54
18	35	35	35	0.43
19	36	36	36	0.34
20	37	37	37	0.46
21	38	38	38	0.22
22	39	39	39	0.26
23	40	40	40	0.38
24	41	42	42	0.42
25	43	43	43	0.50
26	46	45	45	0.42
27	47	47	47	0.59
28	49	49	49	0.53
29	50	51	50	0.26
30	52	52	52	0.30
31	54	54	54	0.44
32	56	56	56	0.37
33	59	59	59	0.99
34	62	63	63	0.68
35	71	70	70	0.18

Appendix Table 12: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Peds PCF Short Form and PROMIS Cognitive Function – Note: Table 10 is recommended.

Peds PCF Short Form Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing	Equipercentile Equivalents with Postsmoothing (More Smoothing)
7	14	14	14
8	20	15	16
9	20	17	18
10	21	19	19
11	22	21	22
12	23	24	25
13	26	27	27
14	30	30	29
15	32	32	31
16	33	33	32
17	34	34	34
18	35	35	35
19	36	36	36
20	37	37	37
21	38	38	38
22	39	39	39
23	40	40	41
24	41	42	42
25	43	43	44
26	46	45	45
27	47	47	47
28	49	49	49
29	51	51	50
30	52	52	52
31	54	54	54
32	56	56	56
33	58	59	59
34	62	62	62
35	69	69	69

Appendix Table 13: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for HADS Anxiety to PROMIS Anxiety (PROsetta Stone Wave 1 study) - RECOMMENDED

HADS Score	PROMIS T-score	SE
0	33.6	6.5
1	37.7	6.1
2	41.1	5.8
3	43.8	5.7
4	46.4	5.5
5	48.7	5.4
6	50.9	5.3
7	52.9	5.2
8	54.9	5.1
9	56.8	5.1
10	58.7	5.1
11	60.5	5.0
12	62.4	5.0
13	64.2	5.0
14	66.1	5.0
15	68.0	5.0
16	70.0	5.0
17	72.0	5.0
18	74.2	5.0
19	76.5	5.0
20	78.9	4.9
21	81.5	4.5

Appendix Table 14: Direct (Raw to Scale) Equipercentile Crosswalk Table – From HADS Anxiety to PROMIS Anxiety – Note: Table 13 is recommended.

HADS Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	34	27	27	0.06
1	34	35	35	0.06
2	42	41	41	0.34
3	45	45	45	0.53
4	48	48	47	0.38
5	50	49	49	0.24
6	51	51	51	0.70
7	54	53	53	0.26
8	55	55	55	0.37
9	56	56	56	0.59
10	58	58	58	0.42
11	61	60	60	0.44
12	62	62	62	0.28
13	64	64	64	0.40
14	66	66	66	1.17
15	68	69	69	2.66
16	73	73	72	1.14
17	75	76	76	0.78
18	82	80	80	1.41
19	84	84	84	0.67
20	85	87	86	0.43
21	85	89	89	0.18

Appendix Table 15: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From HADS Anxiety to PROMIS Anxiety – Note: Table 13 is recommended.

HADS Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	34	34	34
1	36	36	36
2	41	42	42
3	45	45	45
4	48	47	47
5	49	49	49
6	51	51	51
7	53	53	53
8	54	55	55
9	56	56	56
10	58	58	58
11	60	60	60
12	62	62	62
13	64	64	64
14	66	66	67
15	69	69	69
16	73	72	72
17	76	76	75
18	80	79	79
19	84	84	84
20	85	85	84
21	85	85	85

Appendix Table 16: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PANAS Negative Affect to PROMIS Anxiety (PROsetta Stone Wave 1 Study) - RECOMMENDED

PANAS Negative Affect Score	PROMIS T-score	SE	PANAS Negative Affect Score	PROMIS T-score	SE
10	37.4	5.9	40	72.6	2.3
11	43.0	4.2	41	73.5	2.4
12	46.0	3.7	42	74.5	2.4
13	48.1	3.3	43	75.5	2.4
14	49.9	3.0	44	76.5	2.5
15	51.3	2.8	45	77.6	2.5
16	52.6	2.7	46	78.8	2.6
17	53.8	2.6	47	80.2	2.7
18	54.8	2.5	48	81.7	2.8
19	55.8	2.4	49	83.4	2.9
20	56.7	2.4	50	85.1	2.8
21	57.6	2.4			
22	58.5	2.3			
23	59.3	2.3			
24	60.1	2.3			
25	60.9	2.3			
26	61.7	2.3			
27	62.4	2.3			
28	63.2	2.3			
29	63.9	2.3			
30	64.7	2.3			
31	65.5	2.3			
32	66.2	2.3			
33	67.0	2.3			
34	67.7	2.3			
35	68.5	2.3			
36	69.3	2.3			
37	70.1	2.3			
37	70.1	2.3			
38	70.9	2.3			
39	71.8	2.3			

Appendix Table 17: Direct (Raw to Scale) Equipercentile Crosswalk Table – From PANAS Negative Affect to PROMIS Anxiety – Note: Table 16 is recommended.

PANAS Negative Affect Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
10	34	35	35	0.34
11	43	43	43	0.34
12	47	47	47	0.53
13	49	49	49	0.38
14	50	50	50	0.17
15	52	52	52	0.31
16	54	53	53	0.21
17	54	54	54	0.18
18	55	55	55	0.25
19	56	56	56	0.46
20	57	57	57	0.33
21	58	58	58	0.29
22	60	60	59	0.35
23	61	60	60	0.30
24	61	61	61	0.27
25	62	61	61	0.20
26	62	62	62	0.18
27	62	62	62	0.19
28	62	63	63	0.21
29	63	63	64	0.67
30	64	64	64	0.31
31	65	65	65	0.34
32	67	66	66	0.31
33	67	67	67	0.26
34	68	68	68	0.58
35	69	69	69	1.17
36	70	70	69	0.46
37	70	70	70	0.44
38	71	71	71	0.42
39	71	72	72	0.36
40	73	73	73	0.95
41	75	74	74	0.59
42	75	76	75	0.48
43	77	77	77	1.15
44	78	78	78	1.19
45	78	79	79	0.86
46	82	80	80	0.86
47	82	81	81	0.72
48	82	82	82	0.72
49	82	82	83	0.67
50	85	88	88	0.38

Appendix Table 18: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – rom PANAS Negative Affect to PROMIS Anxiety – Note: Table 16 is recommended.

PANAS Negative Affect Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
10	36	37	38
11	44	44	44
12	47	47	47
13	49	49	49
14	50	50	50
15	52	52	52
16	53	53	53
17	54	54	54
18	55	55	55
19	56	56	56
20	57	57	57
21	58	58	58
22	60	59	59
23	60	60	60
24	61	61	61
25	62	62	61
26	62	62	62
27	62	62	62
28	63	63	63
29	63	64	64
30	64	64	64
31	65	65	65
32	66	66	66
33	67	67	67
34	68	68	68
35	69	69	69
36	70	70	70
37	70	70	70
38	71	71	71
39	72	72	72
40	73	73	73
41	74	74	74
42	76	75	75
43	76	76	76
44	78	77	77
45	78	78	78
46	79	79	79
47	82	80	80
48	82	81	81
49	82	82	83
50	85	85	85

Appendix Table 19: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for BDI-II to PROMIS Depression (PROsetta Wave 1 Study) - RECOMMENDED

BDI-II Score	PROMIS T-score	SE	BDI-II Score	PROMIS T-score	SE
0	34.9	5.8	37	68.4	1.8
1	39.4	4.6	38	68.9	1.8
2	42.3	4.0	39	69.4	1.8
3	44.4	3.6	40	69.9	1.8
4	46.2	3.2	41	70.4	1.8
5	47.6	2.9	42	70.9	1.8
6	48.9	2.7	43	71.4	1.8
7	50.0	2.5	44	71.9	1.8
8	51.0	2.4	45	72.4	1.9
9	51.9	2.3	46	72.9	1.9
10	52.7	2.2	47	73.5	1.9
11	53.5	2.1	48	74.0	1.9
12	54.2	2.1	49	74.6	1.9
13	54.9	2.0	50	75.2	1.9
14	55.6	2.0	51	75.7	2.0
15	56.3	2.0	52	76.4	2.0
16	56.9	2.0	53	77.0	2.0
17	57.5	2.0	54	77.7	2.1
18	58.2	2.0	55	78.4	2.2
19	58.8	1.9	56	79.1	2.2
20	59.3	1.9	57	79.9	2.3
21	59.9	1.9	58	80.8	2.4
22	60.5	1.9	59	81.8	2.5
23	61.1	1.9	60	82.9	2.6
24	61.6	1.9	61	84.0	2.6
25	62.2	1.9	62	85.1	2.6
26	62.7	1.9	63	86.3	2.4
27	63.2	1.9			
28	63.8	1.9			
29	64.3	1.9			
30	64.8	1.9			
31	65.3	1.9			
32	65.8	1.9			
33	66.4	1.9			
34	66.9	1.9			
35	67.4	1.8			
36	67.9	1.8			

Appendix Table 20: Direct (Raw to Scale) Equipercentile Crosswalk Table – From BDI-II to PROMIS Depression – Note: Table 19 is recommended.

BDI-II Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	35	33	36	0.20
1	40	40	39	0.20
2	43	43	42	0.21
3	44	45	44	0.31
4	46	46	46	0.28
5	48	48	48	0.57
6	49	49	49	0.25
7	50	50	50	0.71
8	52	51	51	0.34
9	52	52	52	0.30
10	53	53	53	0.31
11	54	54	54	0.22
12	54	54	54	0.20
13	55	55	55	0.34
14	56	56	56	0.35
15	56	56	56	0.33
16	57	57	57	0.38
17	57	57	57	0.36
18	58	58	58	0.35
19	58	58	59	0.35
20	59	59	59	0.33
21	60	60	60	0.38
22	60	60	60	0.34
23	61	61	61	0.54
24	62	61	62	0.23
25	62	62	62	0.22
26	62	62	63	0.22
27	63	63	63	0.31
28	63	64	64	0.30
29	64	64	64	0.61
30	65	65	65	0.52
31	65	65	65	0.51
32	66	66	66	0.58
33	67	67	67	0.34
34	67	67	67	0.31
35	68	68	68	1.44
36	69	68	68	0.65
37	69	69	69	0.61
38	70	70	69	0.35

39	70	70	70	0.33
40	70	70	70	0.31
41	71	71	71	1.92
42	72	71	71	0.47
43	72	72	72	0.45
44	72	72	72	0.39
45	72	73	73	0.38
46	73	73	73	0.49
47	73	73	73	0.44
48	73	74	74	0.42
49	74	74	74	0.80
50	74	74	75	0.73
51	74	75	75	0.70
52	75	75	76	0.87
53	75	76	76	0.71
54	77	76	77	2.45
55	78	77	77	2.45
56	78	77	78	2.45
57	78	78	78	2.45
58	78	79	79	1.00
59	79	79	79	1.00
60	79	80	80	1.06
61	83	83	83	0.41
62	83	86	86	0.31
63	83	89	89	0.15

Appendix Table 21: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From BDI-II to PROMIS Depression – Note: Table 19 is recommended.

BDI-II Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	35	35	35
1	40	39	38
2	42	43	43
3	44	45	45
4	46	46	47
5	48	48	48
6	49	49	49
7	50	50	50
8	51	51	51
9	52	52	52
10	53	53	53
11	54	54	53
12	54	54	54
13	55	55	55
14	56	56	55
15	56	56	56
16	57	57	57
17	57	58	57
18	58	58	58
19	58	59	58
20	59	59	59
21	60	60	60
22	60	60	60
23	61	61	61
24	61	61	61
25	62	62	62
26	62	62	63
27	63	63	63
28	63	64	64
29	64	64	64
30	65	65	65
31	65	65	65
32	66	66	66
33	67	67	66
34	67	67	67
35	68	68	67

36	69	68	68
37	69	69	68
38	70	69	69
39	70	70	70
40	70	70	70
41	71	71	70
42	71	71	71
43	72	72	72
44	72	72	72
45	72	72	72
46	73	73	73
47	73	73	73
48	74	74	74
49	74	74	74
50	74	74	75
51	74	75	75
52	75	75	76
53	75	76	77
54	77	76	77
55	78	77	78
56	78	77	79
57	78	78	80
58	79	79	80
59	79	80	81
60	80	80	81
61	83	81	82
62	83	83	83
63	83	83	83

Appendix Table 22: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for K6 and PROMIS Depression (NIHToolbox Study) - RECOMMENDED

K6 Score	PROMIS T-score	SE
6	36.8	6.7
7	41.5	5.9
8	44.7	5.6
9	47.4	5.3
10	49.5	5.3
11	51.6	5.0
12	53.5	4.8
13	55.2	4.6
14	56.9	4.5
15	58.4	4.4
16	59.9	4.3
17	61.3	4.2
18	62.6	4.1
19	64.0	4.1
20	65.3	4.1
21	66.6	4.1
22	68.0	4.1
23	69.3	4.1
24	70.7	4.1
25	72.2	4.2
26	73.7	4.2
27	75.3	4.3
28	77.1	4.3
29	78.9	4.3
30	81.3	4.2

Appendix Table 23: Direct (Raw to Scale) Equipercentile Crosswalk Table – From K6 to PROMIS Depression – Note: Table 22 is recommended.

K6 Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
6	34	34	35	0.41
7	43	42	41	0.41
8	46	46	46	0.90
9	48	49	49	0.55
10	51	51	51	0.88
11	53	53	53	0.30
12	54	54	54	1.01
13	56	56	56	0.45
14	57	57	57	0.44
15	58	58	58	0.42
16	59	59	60	0.39
17	61	61	61	0.67
18	62	62	62	0.53
19	64	64	64	0.57
20	65	65	65	0.43
21	65	66	66	0.40
22	67	67	68	1.72
23	69	69	69	1.90
24	71	70	70	0.82
25	71	71	72	0.67
26	73	73	73	0.82
27	73	74	74	0.75
28	75	75	76	1.05
29	77	77	77	1.87
30	79	79	79	1.31

Appendix Table 24: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From K6 to PROMIS Depression – Note: Table 22 is recommended.

K6 Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
6	35	34	30
7	42	42	43
8	46	46	46
9	48	48	49
10	51	51	51
11	53	53	52
12	54	54	54
13	56	56	55
14	57	57	57
15	58	58	58
16	59	60	59
17	60	61	61
18	62	62	62
19	64	63	63
20	65	65	65
21	65	66	66
22	67	67	67
23	69	69	69
24	71	70	70
25	71	71	71
26	73	73	73
27	73	74	74
28	76	76	76
29	77	77	78
30	79	79	81

Appendix Table 25: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PANAS Negative Affect and PROMIS Depression (PROsetta Stone Study) – RECOMMENDED

PANAS Negative Affect Score	PROMIS T-score	SE	PANAS Negative Affect Score	PROMIS T-score	SE
10	36.4	5.8	36	68.7	2.5
11	41.7	4.3	37	69.5	2.5
12	44.6	3.8	38	70.4	2.5
13	46.8	3.5	39	71.3	2.6
14	48.5	3.3	40	72.1	2.6
15	50.0	3.1	41	73.1	2.6
16	51.4	3.0	42	74.0	2.6
17	52.6	2.9	43	75.0	2.6
18	53.7	2.8	44	76.1	2.7
19	54.7	2.7	45	77.3	2.7
20	55.7	2.7	46	78.5	2.8
21	56.6	2.6	47	79.9	2.9
22	57.5	2.6	48	81.4	3.0
23	58.4	2.6	49	83.1	3.0
24	59.2	2.6	50	84.9	2.9
25	60.0	2.5			
26	60.8	2.5			
27	61.6	2.5			
28	62.4	2.5			
29	63.2	2.5			
30	64.0	2.5			
31	64.7	2.5			
32	65.5	2.5			
33	66.3	2.5			
34	67.1	2.5			
35	67.9	2.5			

Appendix Table 26: Direct (Raw to Scale) Equipercentile Crosswalk Table – From PANAS Negative Affect to PROMIS Depression – Note: Table 25 is recommended.

PANAS Negative Affect Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
10	35	35	35	0.23
11	43	42	42	0.23
12	46	46	45	0.31
13	47	47	48	0.41
14	49	49	49	0.25
15	51	51	51	0.71
16	52	52	52	0.33
17	54	54	53	0.27
18	54	54	55	0.24
19	55	55	55	0.42
20	56	57	56	0.42
21	58	58	57	0.41
22	58	58	58	0.37
23	59	59	59	0.34
24	60	60	60	0.37
25	60	60	60	0.35
26	61	61	61	0.54
27	62	62	62	0.23
28	62	62	62	0.23
29	62	63	63	0.23
30	63	63	64	0.34
31	65	64	64	0.53
32	66	66	65	0.62
33	67	66	66	0.34
34	67	67	67	0.31
35	68	68	68	1.32
36	69	69	69	0.60
37	70	70	69	0.35
38	70	70	70	0.34
39	71	71	71	2.61
40	72	72	72	0.52
41	73	73	72	0.51
42	73	73	73	0.46
43	74	74	74	0.89
44	75	74	75	1.04
45	75	75	75	0.83
46	75	76	76	0.83
47	76	76	77	2.83
48	77	77	77	2.83
49	78	78	78	2.45
50	83	86	86	0.47

Appendix Table 27: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From PANAS Negative Affect to PROMIS Depression – Note: Table 25 is recommended.

PANAS Negative Affect Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
10	36	36	36
11	42	42	42
12	45	45	46
13	47	47	48
14	49	49	49
15	51	51	51
16	52	52	52
17	54	54	53
18	54	54	54
19	55	56	56
20	57	57	57
21	58	58	57
22	59	59	58
23	59	59	59
24	60	60	60
25	61	60	60
26	61	61	61
27	61	61	62
28	62	62	62
29	62	62	63
30	63	63	64
31	65	64	64
32	66	66	65
33	67	66	66
34	67	67	67
35	68	68	68
36	69	69	69
37	70	70	69
38	70	70	70
39	71	71	71
40	72	72	72
41	73	72	72
42	74	73	73
43	74	74	74
44	74	74	74
45	75	75	75
46	75	76	76
47	77	76	77
48	77	77	78
49	78	78	79
50	83	82	82

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

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

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

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

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

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

Appendix Table 31: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QoL Fatigue and PROMIS Fatigue (PROsetta Stone Wave 1 Study) - RECOMMENDED

Neuro-QoL Fatigue T-Score	Neuro-QoL Fatigue Raw	PROMIS T-score	SE
28.1	19	30.7	4.8
32.1	20	35.3	3.3
33.9	21	37.7	2.8
35.4	22	39.4	2.5
36.5	23	40.8	2.3
37.5	24	42	2.1
38.3	25	43	2
39.0	26	43.9	1.9
39.7	27	44.7	1.8
40.3	28	45.5	1.8
40.9	29	46.2	1.7
41.5	30	46.8	1.7
42.0	31	47.4	1.6
42.5	32	48	1.6
43.0	33	48.6	1.6
43.5	34	49.1	1.5
43.9	35	49.6	1.5
44.4	36	50.2	1.5
44.8	37	50.7	1.5
45.3	38	51.1	1.5
45.7	39	51.6	1.5
46.1	40	52.1	1.5
46.5	41	52.5	1.5
47.0	42	53	1.4
47.4	43	53.4	1.4
47.8	44	53.9	1.4
48.2	45	54.3	1.4
48.7	46	54.7	1.4
49.1	47	55.2	1.4
49.5	48	55.6	1.4
50.0	49	56	1.4
50.4	50	56.5	1.4
50.8	51	56.9	1.4
51.3	52	57.3	1.4
51.7	53	57.8	1.4
52.2	54	58.2	1.4
52.6	55	58.6	1.4
53.1	56	59.1	1.4
53.5	57	59.5	1.4

54.0	58	59.9	1.4
54.4	59	60.4	1.4
54.9	60	60.8	1.5
55.3	61	61.2	1.5
55.8	62	61.7	1.5
56.2	63	62.1	1.5
56.7	64	62.6	1.5
57.1	65	63	1.5
57.6	66	63.5	1.5
58.1	67	63.9	1.5
58.6	68	64.4	1.5
59.0	69	64.9	1.5
59.5	70	65.3	1.5
60.0	71	65.8	1.5
60.5	72	66.3	1.5
61.0	73	66.8	1.5
61.5	74	67.2	1.5
62.0	75	67.7	1.5
62.5	76	68.2	1.5
63.0	77	68.7	1.5
63.5	78	69.2	1.5
64.0	79	69.7	1.5
64.5	80	70.2	1.5
65.1	81	70.8	1.5
65.6	82	71.3	1.5
66.2	83	71.8	1.5
66.8	84	72.4	1.5
67.4	85	73	1.5
68.0	86	73.5	1.6
68.7	87	74.2	1.6
69.4	88	74.8	1.7
70.2	89	75.5	1.7
71.1	90	76.3	1.8
72.1	91	77.2	1.9
73.2	92	78.2	2.1
74.7	93	79.5	2.4
76.6	94	81.1	2.7
79.5	95	83.3	3

Appendix Table 32: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Fatigue to PROMIS Fatigue – Note: Table 31 is recommended.

Neuro-QoL Fatigue T-Score	Neuro-QoL Fatigue Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
28.1	19	30	28	27	0.24
32.1	20	35	34	34	0.24
33.9	21	38	37	37	0.21
35.4	22	38	39	39	0.21
36.5	23	40	40	41	0.21
37.5	24	42	42	42	0.28
38.3	25	43	43	43	0.33
39.0	26	44	44	44	0.45
39.7	27	45	45	45	0.28
40.3	28	46	46	46	0.48
40.9	29	47	47	47	0.38
41.5	30	48	48	47	0.15
42.0	31	48	48	48	0.14
42.5	32	48	48	48	0.14
43.0	33	49	49	49	0.37
43.5	34	49	49	49	0.34
43.9	35	50	50	50	0.19
44.4	36	50	50	50	0.18
44.8	37	50	51	51	0.18
45.3	38	51	51	51	0.37
45.7	39	52	52	52	0.2
46.1	40	52	52	52	0.19
46.5	41	52	53	53	0.19
47.0	42	53	53	53	0.38
47.4	43	54	53	53	0.25
47.8	44	54	54	54	0.23
48.2	45	54	54	54	0.22
48.7	46	54	55	55	0.24
49.1	47	55	55	55	0.48
49.5	48	56	56	56	0.33
50.0	49	57	56	56	0.22
50.4	50	57	57	57	0.2
50.8	51	57	57	57	0.2
51.3	52	57	57	57	0.2
51.7	53	58	58	58	0.41
52.2	54	58	58	58	0.37

52.6	55	59	59	59	0.17
53.1	56	59	59	59	0.17
53.5	57	60	60	60	0.24
54.0	58	60	60	60	0.21
54.4	59	61	61	61	0.44
54.9	60	61	61	61	0.4
55.3	61	62	62	62	0.21
55.8	62	62	62	62	0.21
56.2	63	62	62	62	0.21
56.7	64	62	63	63	0.33
57.1	65	63	63	63	0.3
57.6	66	64	64	64	0.6
58.1	67	64	64	64	0.57
58.6	68	65	65	65	0.21
59.0	69	65	65	65	0.2
59.5	70	66	65	65	0.26
60.0	71	66	66	66	0.23
60.5	72	66	66	66	0.22
61.0	73	66	67	67	0.21
61.5	74	67	67	67	0.43
62.0	75	68	68	68	0.44
62.5	76	68	68	68	0.35
63.0	77	69	69	69	1.28
63.5	78	70	70	70	1.17
64.0	79	71	71	70	0.65
64.5	80	71	71	71	0.52
65.1	81	72	72	72	1.25
65.6	82	74	73	73	0.53
66.2	83	74	74	73	0.39
66.8	84	74	74	74	0.41
67.4	85	74	74	74	0.41
68.0	86	74	75	75	0.94
68.7	87	75	75	76	0.94
69.4	88	75	76	76	0.9
70.2	89	75	77	77	0.9
71.1	90	76	78	79	1.41
72.1	91	78	79	80	4
73.2	92	83	83	82	0.26
74.7	93	83	84	83	0.26
76.6	94	83	87	86	0.26
79.5	95	83	89	89	0.18

Appendix Table 33: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Fatigue to PROMIS Fatigue – Note: Table 31 is recommended.

Neuro-QoL Fatigue T-Score	Neuro-QoL Fatigue Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
28.1	19	30	30	30
32.1	20	34	33	33
33.9	21	37	37	37
35.4	22	39	39	39
36.5	23	40	40	41
37.5	24	42	42	42
38.3	25	43	43	43
39.0	26	44	44	44
39.7	27	45	45	45
40.3	28	46	46	46
40.9	29	47	47	46
41.5	30	48	47	47
42.0	31	48	48	48
42.5	32	48	48	48
43.0	33	49	49	49
43.5	34	49	49	49
43.9	35	50	50	50
44.4	36	50	50	50
44.8	37	51	51	51
45.3	38	51	51	51
45.7	39	52	52	52
46.1	40	52	52	52
46.5	41	53	53	53
47.0	42	53	53	53
47.4	43	54	54	54
47.8	44	54	54	54
48.2	45	54	54	54
48.7	46	55	55	55
49.1	47	55	55	55
49.5	48	56	56	56
50.0	49	57	56	56
50.4	50	57	57	57
50.8	51	57	57	57
51.3	52	57	57	57
51.7	53	58	58	58
52.2	54	58	58	58
52.6	55	58	59	59

53.1	56	59	59	59
53.5	57	60	60	60
54.0	58	60	60	60
54.4	59	61	61	60
54.9	60	61	61	61
55.3	61	62	62	61
55.8	62	62	62	62
56.2	63	62	62	62
56.7	64	62	63	63
57.1	65	63	63	63
57.6	66	64	64	64
58.1	67	64	64	64
58.6	68	65	65	65
59.0	69	65	65	65
59.5	70	66	66	66
60.0	71	66	66	66
60.5	72	66	66	66
61.0	73	67	67	67
61.5	74	67	67	68
62.0	75	68	68	68
62.5	76	68	68	69
63.0	77	69	69	69
63.5	78	70	70	70
64.0	79	71	71	70
64.5	80	71	71	71
65.1	81	72	72	71
65.6	82	73	73	72
66.2	83	74	73	73
66.8	84	74	74	73
67.4	85	75	74	74
68.0	86	75	75	75
68.7	87	75	76	76
69.4	88	75	76	76
70.2	89	76	77	77
71.1	90	76	78	79
72.1	91	78	79	80
73.2	92	82	81	81
74.7	93	82	82	82
76.6	94	83	83	83
79.5	95	83	83	83

Appendix Table 34: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for VR-12 –Mental and PROMIS Global Health - Mental (PROsetta Stone Wave 2 Study) - RECOMMENDED

VR-12 – Mental Raw Score	PROMIS T-score	SE
6	17.9	4.3
7	20.4	4.5
8	22.7	4.5
9	24.9	4.5
10	26.8	4.5
11	28.6	4.5
12	30.3	4.4
13	31.9	4.4
14	33.4	4.4
15	34.9	4.4
16	36.4	4.4
17	37.8	4.4
18	39.2	4.4
19	40.6	4.4
20	42.0	4.4
21	43.4	4.4
22	44.9	4.4
23	46.4	4.5
24	47.9	4.5
25	49.5	4.6
26	51.1	4.6
27	52.9	4.7
28	54.7	4.8
29	56.7	4.9
30	59.0	5.0
31	61.8	5.1
32	65.4	5.4
33	70.2	6.0

Appendix Table 35: Direct (Raw to Scale) Equipercentile Crosswalk Table – From VR-12 –Mental to PROMIS Global Health - Mental – Note: Table 34 is recommended.

VR-12 Mental Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
6	21	12	12	0.09
7	21	16	16	0.09
8	21	20	20	0.19
9	25	25	25	0.33
10	25	26	26	0.37
11	28	28	28	0.17
12	28	29	29	0.21
13	31	31	31	0.19
14	31	32	32	0.22
15	34	34	34	0.17
16	36	35	35	0.19
17	36	37	37	0.19
18	39	39	39	0.11
19	41	40	40	0.13
20	41	42	42	0.13
21	44	44	44	0.10
22	46	45	45	0.10
23	46	46	46	0.10
24	48	48	48	0.11
25	51	50	50	0.11
26	51	51	51	0.09
27	53	52	52	0.10
28	53	53	53	0.09
29	56	56	55	0.12
30	59	58	58	0.20
31	62	62	62	0.16
32	68	68	69	0.12
33	68	75	76	0.07

Appendix Table 36: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From VR-12 –Mental to PROMIS Global Health - Mental – Note: Table 34 is recommended.

VR-12 Mental Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
6	20	20	20
7	20	22	22
8	23	23	23
9	25	24	24
10	26	26	26
11	28	28	28
12	30	30	30
13	31	31	31
14	32	32	33
15	34	34	34
16	35	35	36
17	37	37	37
18	39	39	39
19	40	40	40
20	42	42	42
21	44	44	43
22	45	45	45
23	46	46	46
24	48	48	48
25	50	49	49
26	51	51	51
27	52	52	52
28	54	54	54
29	55	56	56
30	58	58	58
31	62	62	61
32	66	66	65
33	70	72	71

Appendix Table 37: Direct (Raw-to-Raw-to-Scale, more smoothing) Equipercentile Crosswalk Table – From Algorithmic VR-12 – Mental Health Component to PROMIS Global Health - Mental (PROsetta Stone Wave 2 Study)

VR-12 Mental Score	PROMIS T-score	VR-12 Mental Score	PROMIS T-score
9	20.5	39	42.3
10	21.3	40	43.1
11	22.1	41	43.8
12	22.9	42	44.4
13	23.6	43	45.1
14	24.3	44	45.8
15	25.1	45	46.5
16	25.8	46	47.3
17	26.6	47	48.0
18	27.4	48	48.8
19	28.2	49	49.5
20	28.9	50	50.3
21	29.7	51	51.0
22	30.4	52	51.8
23	31.1	53	52.6
24	31.8	54	53.5
25	32.5	55	54.4
26	33.1	56	55.4
27	33.8	57	56.5
28	34.5	58	57.7

29	35.2	59	59.0
30	35.9	60	60.3
31	36.6	61	61.9
32	37.4	62	63.7
33	38.1	63	65.9
34	38.8	64	68.4
35	39.5	65	70.6
36	40.2	66	71.0
37	40.9	67	71.0
38	41.6	68	71.1

Appendix Table 38: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for VR-12 – Physical to PROMIS Global Health - Physical (PROsetta Stone Wave 2 Study) - RECOMMENDED

VR-12 – Physical Raw Score	PROMIS T-score	SE
7	19.4	4.2
8	22.2	4.1
9	24.4	3.9
10	26.4	3.8
11	28.1	3.7
12	29.7	3.7
13	31.2	3.6
14	32.7	3.6
15	34.1	3.6
16	35.5	3.6
17	36.9	3.6
18	38.2	3.6
19	39.6	3.6
20	41.0	3.6
21	42.4	3.7
22	43.9	3.7
23	45.4	3.8
24	46.9	3.8
25	48.6	4.0
26	50.4	4.1
27	52.3	4.3
28	54.4	4.5
29	56.8	4.7
30	59.6	4.9
31	63.4	5.5
32	67.8	6.2

Appendix Table 39: Direct (Raw to Scale) Equipercentile Crosswalk Table – From VR-12 – Physical to PROMIS Global Health - Physical – Note: Table 38 is recommended.

VR-12 Physical Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
7	18	13	13	0.43
8	24	20	20	0.68
9	27	26	26	0.16
10	27	27	27	0.17
11	27	28	28	0.21
12	30	30	30	0.19
13	32	31	31	0.16
14	32	32	32	0.15
15	35	34	34	0.11
16	35	35	35	0.11
17	35	36	36	0.12
18	37	37	37	0.10
19	40	39	39	0.10
20	40	41	41	0.09
21	42	42	42	0.09
22	45	45	44	0.09
23	45	46	46	0.08
24	48	48	48	0.09
25	48	49	49	0.09
26	51	51	51	0.06
27	51	52	52	0.06
28	54	54	54	0.06
29	54	55	55	0.06
30	58	58	58	0.08
31	62	62	62	0.10
32	68	73	74	0.09

Appendix Table 40: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From VR-12 – Physical to PROMIS Global Health - Physical – Note: Table 38 is recommended.

VR-12 Physical Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
7	19	17	17
8	23	21	21
9	26	25	24
10	27	27	27
11	28	28	28
12	30	30	30
13	31	31	31
14	32	32	32
15	34	33	33
16	35	35	35
17	36	36	36
18	37	38	38
19	39	39	39
20	41	41	41
21	43	42	42
22	44	44	44
23	46	46	46
24	48	47	47
25	49	49	49
26	51	51	51
27	52	52	52
28	54	54	54
29	56	56	56
30	59	59	59
31	63	63	63
32	69	69	68

Appendix Table 41: Direct (Raw-to-Raw-to-Scale, More Smoothing) Equipercentile Crosswalk Table – From Algorithmic VR-12 – Physical Health Component to PROMIS Global Health - Physical (PROsetta Stone Wave 2 Study)

VR-12 Physical Score	PROMIS T-score	VR-12 Physical Score	PROMIS T-score
10	16.6	41	43.1
11	17.4	42	43.9
12	18.4	43	44.7
13	19.6	44	45.5
14	20.9	45	46.3
15*	21.6	46	47.2
16	22.2	47	48.0
17	23.6	48	48.8
18	24.6	49	49.7
19	25.5	50	50.6
20	26.4	51	51.5
21	27.3	52	52.5
22	28.1	53	53.5
23	28.9	54	54.7
24	29.8	55	56.0
25	30.6	56	57.5
26	31.5	57	59.1
27	32.3	58	60.8
28	33.1	59	62.7

29	33.9	60	64.8
30	34.6	61	67.2
31	35.4	62	69.8
32	36.1	63	71.2
33	36.9	64	71.4
34	37.7	65	71.5
35	38.4		
36	39.2		
37	40.0		
38	40.7		
39	41.5		
40	42.3		

*No participant scored 15; we linearly interpolated the PROMIS T-score.

Appendix Table 42: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for SF-36 /BP and PROMIS Pain Interference (PROsetta Stone Wave 1 Study) - RECOMMENDED

SF-36 /BP Score	PROMIS T-score	SE
2	37.8	6.2
3	45.1	4.5
4	49.9	4.1
5	53.3	3.9
6	56.6	3.7
7	60.1	3.6
8	63.5	3.7
9	67.0	3.7
10	71.1	4.0
11	76.0	4.6

Appendix Table 43: Direct (Raw to Scale) Equipercentile Crosswalk Table – From SF-36 /BP to PROMIS Pain Interference – Note: Table 42 is recommended.

SF-36 /BP 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	37	33	32	0.65
3	45	45	45	0.65
4	51	50	50	0.56
5	54	54	54	0.47
6	57	57	57	0.76
7	60	60	60	0.99
8	63	64	64	0.89
9	67	67	67	0.83
10	72	71	71	3.16
11	76	84	84	0.47

Appendix Table 44: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From SF-36 /BP to PROMIS Pain Interference – Note: Table 42 is recommended.

SF-36 /BP Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
2	36	25	-54
3	45	45	46
4	51	50	51
5	54	54	54
6	57	57	57
7	60	60	60
8	63	63	63
9	67	67	67
10	72	72	71
11	75	77	76

Appendix Table 45: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QoL Sleep Disturbance and PROMIS Sleep Disturbance (PROsetta Stone Wave 2 Study) - RECOMMENDED

Neuro-QoL Sleep Disturbance T-Score	Neuro-QoL Sleep Disturbance Raw Score	PROMIS T-score	SE
32.0	8	30.7	6.0
36.3	9	34.5	5.3
39.1	10	37.3	5.0
41.7	11	39.8	4.7
43.8	12	41.9	4.5
45.6	13	43.8	4.3
47.3	14	45.6	4.1
48.9	15	47.2	3.9
50.4	16	48.7	3.8
51.8	17	50.1	3.8
53.1	18	51.5	3.7
54.4	19	52.8	3.6
55.6	20	54.0	3.6
56.8	21	55.2	3.6
58.0	22	56.4	3.6
59.2	23	57.6	3.6
60.4	24	58.8	3.5
61.6	25	59.9	3.5
62.8	26	61.1	3.5
63.9	27	62.2	3.5
65.1	28	63.4	3.5
66.4	29	64.5	3.5
67.6	30	65.7	3.6
68.9	31	67.0	3.6
70.3	32	68.2	3.6
71.7	33	69.5	3.7
73.2	34	70.9	3.7
74.7	35	72.3	3.8
76.4	36	73.8	3.9
78.2	37	75.4	4.0
80.2	38	77.2	4.1
82.2	39	79.1	4.2
84.2	40	81.4	4.2

Appendix Table 46: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Sleep Disturbance to PROMIS Sleep Disturbance – Note: Table 45 is recommended.

Neuro-QoL Sleep Disturbance T-Score	Neuro-QoL Sleep Disturbance Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
32.0	8	31	32	32	3.16
36.3	9	35	35	35	0.56
39.1	10	37	37	38	0.82
41.7	11	40	40	40	0.37
43.8	12	42	42	42	0.34
45.6	13	44	44	44	0.45
47.3	14	46	46	46	0.30
48.9	15	48	48	48	0.47
50.4	16	50	50	50	0.52
51.8	17	52	52	51	0.33
53.1	18	53	53	53	0.59
54.4	19	54	54	54	0.24
55.6	20	56	56	55	0.33
56.8	21	57	57	57	0.28
58.0	22	58	58	58	0.15
59.2	23	58	59	59	0.15
60.4	24	59	60	60	0.25
61.6	25	61	61	61	0.44
62.8	26	62	62	62	0.33
63.9	27	63	63	63	0.44
65.1	28	64	64	64	0.50
66.4	29	66	65	65	0.44
67.6	30	66	66	66	0.36
68.9	31	67	67	67	0.86
70.3	32	68	68	68	1.33
71.7	33	70	69	69	0.93
73.2	34	71	70	70	0.73
74.7	35	71	71	71	0.70
76.4	36	72	71	72	1.37
78.2	37	72	72	72	1.25
80.2	38	73	73	74	0.94
82.2	39	73	74	78	0.82
84.2	40	74	85	86	2.52

Appendix Table 47: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Sleep Disturbance to PROMIS Sleep Disturbance – Note: Table 45 is recommended.

Neuro-QoL Sleep Disturbance T-Score	Neuro-QoL Sleep Disturbance Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
32.0	8	31	30	29
36.3	9	35	35	34
39.1	10	38	38	38
41.7	11	40	40	40
43.8	12	42	42	42
45.6	13	44	44	44
47.3	14	46	46	46
48.9	15	48	48	48
50.4	16	50	50	50
51.8	17	52	52	51
53.1	18	53	53	53
54.4	19	54	54	54
55.6	20	56	56	55
56.8	21	57	57	56
58.0	22	58	58	58
59.2	23	58	58	58
60.4	24	59	59	60
61.6	25	61	60	61
62.8	26	62	62	62
63.9	27	63	63	63
65.1	28	64	64	64
66.4	29	66	65	65
67.6	30	66	66	66
68.9	31	67	67	67
70.3	32	68	68	68
71.7	33	70	69	69
73.2	34	71	70	70
74.7	35	71	71	71
76.4	36	72	72	72
78.2	37	72	72	73
80.2	38	73	73	75
82.2	39	73	74	77
84.2	40	74	80	82

Appendix Table 48: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PROMIS Sleep-related Impairment and PROMIS Sleep Disturbance (PROsetta Stone Wave 2 Study) – RECOMMENDED

PROMIS Sleep-related Impairment T-Score	PROMIS Sleep-related Impairment Raw Score	PROMIS Sleep Disturbance T-score	SE
26.2	16	27.3	5.2
30.0	17	31.0	4.5
32.7	18	33.6	4.1
34.9	19	35.6	3.8
36.8	20	37.3	3.6
38.4	21	38.8	3.4
39.9	22	40.2	3.2
41.2	23	41.4	3
42.4	24	42.5	2.9
43.5	25	43.5	2.8
44.5	26	44.4	2.6
45.5	27	45.3	2.5
46.4	28	46.1	2.4
47.3	29	46.9	2.4
48.1	30	47.6	2.3
48.9	31	48.3	2.2
49.6	32	49.0	2.2
50.3	33	49.6	2.1
51.0	34	50.2	2.1
51.7	35	50.8	2.1
52.3	36	51.4	2.1
53.0	37	52.0	2
53.6	38	52.5	2
54.2	39	53.1	2
54.8	40	53.6	2
55.4	41	54.1	2
55.9	42	54.6	2
56.5	43	55.2	2
57.1	44	55.7	2
57.7	45	56.2	2
58.3	46	56.7	2
58.8	47	57.2	2
59.4	48	57.7	2
60.0	49	58.2	1.9
60.5	50	58.7	1.9

61.1	51	59.2	1.9
61.6	52	59.7	1.9
62.2	53	60.2	1.9
62.7	54	60.7	1.9
63.3	55	61.2	1.9
63.8	56	61.7	1.9
64.4	57	62.2	1.9
64.9	58	62.7	1.9
65.5	59	63.2	1.9
66.1	60	63.7	1.9
66.6	61	64.2	1.9
67.2	62	64.8	2
67.8	63	65.3	2
68.4	64	65.8	2
69.0	65	66.4	2
69.6	66	66.9	2
70.2	67	67.5	2
70.9	68	68.1	2.1
71.6	69	68.7	2.1
72.3	70	69.4	2.1
73.0	71	70.1	2.2
73.8	72	70.8	2.3
74.6	73	71.6	2.4
75.5	74	72.4	2.5
76.4	75	73.3	2.6
77.5	76	74.4	2.8
78.7	77	75.5	3
80.0	78	76.9	3.2
81.5	79	78.5	3.4
83.3	80	80.9	3.7

Appendix Table 49: Direct (Raw to Scale) Equipercentile Crosswalk Table – From PROMIS Sleep-related Impairment to PROMIS Sleep Disturbance – Note: Table 48 is recommended.

PROMIS Sleep-related Impairment T-Score	PROMIS Sleep-related Impairment Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
26.2	16	29	30	31	0.79
30.0	17	34	33	33	0.55
32.7	18	34	35	35	0.40
34.9	19	36	36	36	0.75
36.8	20	38	38	38	0.58
38.4	21	39	39	39	0.76
39.9	22	40	40	40	0.39
41.2	23	42	42	41	0.39
42.4	24	43	43	42	0.46
43.5	25	44	44	43	0.46
44.5	26	44	45	44	0.90
45.5	27	46	46	45	0.32
46.4	28	46	46	46	0.30
47.3	29	47	47	47	0.53
48.1	30	48	48	48	0.50
48.9	31	49	49	48	0.30
49.6	32	49	49	49	0.31
50.3	33	50	50	50	0.53
51.0	34	50	51	50	0.51
51.7	35	51	51	51	0.59
52.3	36	52	52	52	0.34
53.0	37	52	52	52	0.32
53.6	38	53	53	53	0.67
54.2	39	54	53	53	0.29
54.8	40	54	54	54	0.28
55.4	41	54	54	55	0.28
55.9	42	55	55	55	0.47
56.5	43	55	56	56	0.46
57.1	44	56	56	56	0.35
57.7	45	56	57	57	0.35
58.3	46	57	57	57	0.31
58.8	47	58	58	58	0.17
59.4	48	58	58	58	0.17
60.0	49	59	59	59	0.24

60.5	50	59	59	59	0.24
61.1	51	60	60	60	0.30
61.6	52	60	60	60	0.29
62.2	53	61	61	61	0.43
62.7	54	61	61	61	0.39
63.3	55	62	62	62	0.35
63.8	56	62	62	62	0.31
64.4	57	63	63	63	0.43
64.9	58	63	63	63	0.4
65.5	59	64	64	64	0.48
66.1	60	65	64	65	1.08
66.6	61	65	65	65	1.06
67.2	62	66	66	66	0.4
67.8	63	66	66	66	0.35
68.4	64	66	67	67	0.32
69.0	65	66	67	67	0.76
69.6	66	67	68	68	0.74
70.2	67	68	68	68	1.24
70.9	68	69	69	69	0.88
71.6	69	70	69	69	0.81
72.3	70	70	70	70	0.76
73.0	71	71	70	70	0.69
73.8	72	71	71	71	0.61
74.6	73	71	71	71	0.6
75.5	74	72	72	72	1.15
76.4	75	72	72	72	1.1
77.5	76	73	73	73	0.85
78.7	77	73	75	75	0.67
80.0	78	73	77	77	0.67
81.5	79	74	83	83	1.58
83.3	80	76	88	88	0.77

Appendix Table 50: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From PROMIS Sleep-related Impairment to PROMIS Sleep Disturbance – Note: Table 48 is recommended

PROMIS Sleep-related Impairment T-Score	PROMIS Sleep-related Impairment Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
26.2	16	28	29	31
30.0	17	33	32	33
32.7	18	34	35	35
34.9	19	36	36	37
36.8	20	38	38	38
38.4	21	39	39	40
39.9	22	40	40	41
41.2	23	42	42	42
42.4	24	43	43	43
43.5	25	44	44	44
44.5	26	45	45	44
45.5	27	46	46	45
46.4	28	46	46	46
47.3	29	47	47	47
48.1	30	48	48	48
48.9	31	49	49	48
49.6	32	49	49	49
50.3	33	50	50	50
51.0	34	50	51	50
51.7	35	51	51	51
52.3	36	52	52	51
53.0	37	52	52	52
53.6	38	53	53	53
54.2	39	54	53	53
54.8	40	54	54	54
55.4	41	54	54	54
55.9	42	55	55	55
56.5	43	55	56	55
57.1	44	56	56	56
57.7	45	56	56	56
58.3	46	57	57	57
58.8	47	58	58	58
59.4	48	58	58	58
60.0	49	59	59	59

60.5	50	59	59	59
61.1	51	60	60	60
61.6	52	60	60	60
62.2	53	60	61	61
62.7	54	61	61	61
63.3	55	62	62	62
63.8	56	62	62	62
64.4	57	63	63	63
64.9	58	63	63	63
65.5	59	64	64	64
66.1	60	65	64	64
66.6	61	65	65	65
67.2	62	66	65	65
67.8	63	66	66	66
68.4	64	66	66	66
69.0	65	67	67	67
69.6	66	67	68	68
70.2	67	68	68	68
70.9	68	69	69	69
71.6	69	70	69	70
72.3	70	70	70	70
73.0	71	71	70	71
73.8	72	71	71	72
74.6	73	71	72	72
75.5	74	72	72	73
76.4	75	72	73	74
77.5	76	73	73	75
78.7	77	73	75	76
80.0	78	74	76	78
81.5	79	74	79	80
83.3	80	77	84	84

Appendix Table 51: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PSQI to PROMIS Sleep Disturbance (PROMIS Wave 1 Study) - RECOMMENDED

PSQI Score	PROMIS T-score	SE
0	30.4	5.6
1	35.3	5.0
2	39.5	5.0
3	41.0	6.5
4	43.0	6.4
5	45.1	6.0
6	47.2	5.7
7	49.2	5.6
8	51.2	5.6
9	53.1	5.5
10	54.9	5.5
11	56.8	5.4
12	58.5	5.3
13	60.3	5.3
14	62.1	5.3
15	63.8	5.3
16	65.6	5.4
17	67.5	5.4
18	69.4	5.2
19	71.5	4.9
20	74.4	5.0
21	77.6	5.1

Appendix Table 52: Direct (Raw to Scale) Equipercentile Crosswalk Table – From PSQI to PROMIS Sleep Disturbance – Note: Table 51 is recommended.

PSQI Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
0	29	28	29	0.21
1	34	33	33	0.46
2	37	37	37	0.47
3	40	40	40	0.20
4	43	43	43	0.26
5	46	46	46	0.23
6	48	48	48	0.31
7	50	50	50	0.38
8	52	52	52	0.31
9	54	54	54	0.29
10	56	56	56	0.28
11	58	58	58	0.36
12	60	60	60	0.40
13	62	62	62	0.52
14	64	64	64	0.40
15	66	66	66	0.41
16	69	69	69	1.30
17	71	71	71	1.50
18	74	74	73	1.02
19	78	78	78	4.24
20	81	83	83	1.41
21	90	88	88	0.01

Appendix Table 53: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From PSQI to PROMIS Sleep Disturbance - Note: Table 51 is recommended.

PSQI Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	28	28	27
1	33	33	33
2	37	37	37
3	40	40	41
4	43	43	44
5	46	46	46
6	48	48	48
7	50	50	50
8	52	52	52
9	54	54	54
10	56	56	56
11	58	58	58
12	60	60	60
13	62	62	62
14	64	64	64
15	66	66	66
16	69	68	68
17	71	71	71
18	74	74	74
19	78	76	76
20	81	79	79
21	85	84	84

Appendix Table 54: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QoL Sleep Disturbance and PROMIS Sleep-related Impairment (PROsetta Stone Wave 2 Study) - RECOMMENDED

Neuro-QoL Sleep Disturbance T-Score	Neuro-QoL Sleep Disturbance Raw Score	PROMIS T-score	SE
32.0	8	31.0	5.9
36.3	9	35.5	5.3
39.1	10	38.6	5.0
41.7	11	41.1	4.8
43.8	12	43.3	4.6
45.6	13	45.3	4.5
47.3	14	47.1	4.3
48.9	15	48.7	4.2
50.4	16	50.3	4.2
51.8	17	51.8	4.1
53.1	18	53.2	4.0
54.4	19	54.6	4.0
55.6	20	55.9	3.9
56.8	21	57.2	3.9
58.0	22	58.5	3.9
59.2	23	59.8	3.8
60.4	24	61.0	3.8
61.6	25	62.2	3.8
62.8	26	63.4	3.8
63.9	27	64.7	3.8
65.1	28	65.9	3.8
66.4	29	67.1	3.8
67.6	30	68.4	3.8
68.9	31	69.7	3.8
70.3	32	71.0	3.8
71.7	33	72.3	3.9
73.2	34	73.8	3.9
74.7	35	75.2	3.9
76.4	36	76.8	4
78.2	37	78.4	4
80.2	38	80.2	4
82.2	39	82	3.9
84.2	40	83.8	3.6

Appendix Table 55: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Sleep Disturbance to PROMIS Sleep-related Impairment – Note: Table 54 is recommended.

Neuro-QoL Sleep Disturbance T-Score	Neuro-QoL Sleep Disturbance Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
32.0	8	26	27	28	0.40
36.3	9	33	33	33	0.40
39.1	10	37	37	36	0.31
41.7	11	40	39	39	0.39
43.8	12	41	42	42	0.46
45.6	13	44	44	44	0.29
47.3	14	46	46	46	0.23
48.9	15	48	48	48	0.45
50.4	16	50	50	50	0.41
51.8	17	52	52	52	0.38
53.1	18	54	54	54	0.30
54.4	19	55	55	55	0.31
55.6	20	57	57	56	0.77
56.8	21	58	58	58	0.23
58.0	22	59	59	59	0.12
59.2	23	59	59	60	0.13
60.4	24	60	61	61	0.25
61.6	25	62	62	62	0.35
62.8	26	63	63	63	0.32
63.9	27	64	64	64	0.32
65.1	28	66	66	66	0.48
66.4	29	67	67	67	0.72
67.6	30	68	68	68	0.58
68.9	31	69	69	69	2.03
70.3	32	71	71	71	0.65
71.7	33	72	72	72	1.28
73.2	34	73	73	73	1.92
74.7	35	75	74	74	1.04
76.4	36	76	75	75	0.91
78.2	37	76	76	77	1.41
80.2	38	78	77	78	1.06
82.2	39	78	78	79	0.97
84.2	40	82	86	87	1.63

Appendix Table 56: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Sleep Disturbance to PROMIS Sleep-related Impairment – Table 54 is recommended

Neuro-QoL Sleep Disturbance T-Score	Neuro-QoL Sleep Disturbance Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
32.0	8	27	28	27
36.3	9	33	33	33
39.1	10	36	36	36
41.7	11	39	39	40
43.8	12	42	42	42
45.6	13	44	44	44
47.3	14	46	46	46
48.9	15	48	48	48
50.4	16	50	50	50
51.8	17	52	52	52
53.1	18	54	54	54
54.4	19	55	55	55
55.6	20	57	57	56
56.8	21	58	58	58
58.0	22	59	59	59
59.2	23	60	60	60
60.4	24	61	61	61
61.6	25	62	62	62
62.8	26	64	63	63
63.9	27	64	64	64
65.1	28	65	66	66
66.4	29	67	67	67
67.6	30	68	68	68
68.9	31	69	69	69
70.3	32	71	70	70
71.7	33	71	72	71
73.2	34	73	73	72
74.7	35	74	74	73
76.4	36	75	75	75
78.2	37	76	76	76
80.2	38	77	77	78
82.2	39	78	78	79
84.2	40	82	82	82

Appendix Table 57: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PSQI to PROMIS Sleep-related Impairment (PROMIS Wave 1 Study) - RECOMMENDED

PSQI Score	PROMIS T-score	SE
0	30.5	5.9
1	35.4	5.4
2	39.5	5.3
3	42.0	6.0
4	44.1	6.2
5	46.0	6.1
6	48.0	6.0
7	49.9	5.9
8	51.7	5.9
9	53.4	5.9
10	55.2	5.8
11	57.0	5.8
12	58.7	5.7
13	60.5	5.7
14	62.2	5.7
15	64.0	5.6
16	65.9	5.6
17	67.8	5.6
18	69.8	5.5
19	71.9	5.1
20	74.9	5.0
21	78.4	5.0

Appendix Table 58: Direct (Raw to Scale) Equipercentile Crosswalk Table – From PSQI to PROMIS Sleep-related Impairment – Note: Table 57 is recommended.

PSQI 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	30	29	30	0.25
1	35	34	34	0.23
2	38	38	38	0.24
3	41	41	41	0.31
4	44	44	44	0.16
5	46	46	46	0.19
6	48	48	48	0.31
7	50	50	50	0.22
8	52	52	52	0.21
9	54	54	54	0.25
10	56	56	56	0.33
11	58	58	58	0.34
12	59	60	60	0.29
13	62	62	62	0.37
14	63	64	64	0.33
15	66	66	66	0.38
16	68	68	68	0.51
17	70	70	70	0.57
18	73	73	72	1.73
19	76	78	77	1.22
20	83	83	83	1.41
21	90	88	88	0.01

Appendix table 59: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From PSQI to PROMIS Sleep-related Impairment – Note: Table 57 is recommended.

PSQI Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	29	28	26
1	34	34	34
2	38	38	38
3	41	41	41
4	44	44	44
5	46	46	46
6	48	48	48
7	50	50	50
8	52	52	52
9	54	54	54
10	56	56	56
11	58	58	58
12	60	60	60
13	62	62	61
14	63	64	63
15	66	66	65
16	68	68	67
17	70	70	70
18	73	72	72
19	76	75	74
20	83	78	77
21	84	82	81

Appendix Table 60: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PROMIS Satisfaction w/ Participation in Discretionary Social Activities v1.0 and PROMIS Satisfaction w/ Social Roles & Activities v2.0 (PROsetta Stone Wave 2 Study)- RECOMMENDED

PROMIS Satisfaction w/ Participation in DSA v1.0 T-score Score	PROMIS Satisfaction w/ Participation in DSA v1.0 Raw Score	PROMIS Satisfaction w/ Social Roles & Activities v2.0 T-score	SE
26.8	12	22.7	4.1
30.6	13	25.8	3.3
32.4	14	27.7	3.0
33.8	15	29.2	2.7
34.8	16	30.4	2.6
35.8	17	31.5	2.4
36.6	18	32.5	2.3
37.3	19	33.4	2.3
38.0	20	34.2	2.2
38.7	21	35.0	2.2
39.3	22	35.8	2.1
39.9	23	36.5	2.1
40.4	24	37.2	2.1
41.0	25	37.9	2.1
41.6	26	38.6	2.1
42.1	27	39.3	2.1
42.7	28	40.0	2.1
43.2	29	40.6	2.1
43.8	30	41.3	2.1
44.4	31	42.0	2.1
44.9	32	42.6	2.1
45.5	33	43.3	2.1
46.1	34	43.9	2.1
46.6	35	44.6	2.1
47.2	36	45.3	2.1
47.8	37	45.9	2.1
48.4	38	46.6	2.1
49.0	39	47.3	2.1
49.6	40	48.0	2.1
50.2	41	48.7	2.1
50.8	42	49.4	2.1
51.4	43	50.1	2.1
52.0	44	50.9	2.1
52.6	45	51.6	2.1
53.2	46	52.3	2.1
53.8	47	53.1	2.2
54.4	48	53.9	2.2

55.1	49	54.7	2.2
55.7	50	55.5	2.2
56.4	51	56.4	2.2
57.1	52	57.3	2.2
57.8	53	58.2	2.3
58.5	54	59.2	2.4
59.3	55	60.3	2.5
60.3	56	61.5	2.6
61.3	57	63.0	2.8
62.7	58	64.7	3.1
64.6	59	66.9	3.6
68.9	60	70.8	4.7

Appendix Table 61: Direct (Raw to Scale) Equipercentile Crosswalk Table - From PROMIS Satisfaction w/ Participation in Discretionary Social Activities v1.0 to PROMIS Satisfaction w/ Social Roles & Activities v2.0 (PROsetta Stone Wave 2 Study) - Note: Table 60 is recommended.

PROMIS Satisfaction w/ Participation in DSA v1.0 T-score Score	PROMIS Satisfaction w/ Participation in DSA v1.0 Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
26.8	12	20	21	23	0.81
30.6	13	26	26	26	0.81
32.4	14	29	28	28	1.37
33.8	15	31	30	30	0.82
34.8	16	32	32	31	1.54
35.8	17	33	34	33	0.83
36.6	18	35	35	34	1.62
37.3	19	36	36	35	0.49
38.0	20	36	36	36	0.53
38.7	21	37	37	37	0.50
39.3	22	38	38	37	0.29
39.9	23	38	38	38	0.28
40.4	24	39	39	39	0.49
41.0	25	40	39	39	0.40
41.6	26	40	40	40	0.40
42.1	27	41	40	40	0.25
42.7	28	41	41	41	0.24
43.2	29	41	41	41	0.25
43.8	30	42	42	42	0.25
44.4	31	42	42	42	0.25
44.9	32	43	43	43	0.31
45.5	33	43	43	43	0.31
46.1	34	44	44	44	0.13
46.6	35	44	44	44	0.13
47.2	36	45	45	45	0.20
47.8	37	46	46	46	0.25
48.4	38	46	46	46	0.23
49.0	39	47	47	47	0.29
49.6	40	48	48	48	0.24
50.2	41	48	48	48	0.22
50.8	42	49	49	49	0.32
51.4	43	50	50	50	0.28
52.0	44	50	50	50	0.25
52.6	45	51	51	51	0.29
53.2	46	52	52	52	0.14

53.8	47	52	52	52	0.13
54.4	48	53	53	53	0.41
55.1	49	54	54	54	0.27
55.7	50	55	55	55	0.63
56.4	51	56	56	56	0.46
57.1	52	56	57	57	0.41
57.8	53	58	58	58	0.87
58.5	54	59	59	59	0.63
59.3	55	60	60	60	0.57
60.3	56	60	61	62	1.42
61.3	57	62	63	64	0.89
62.7	58	64	65	66	1.21
64.6	59	67	68	68	0.79
68.9	60	71	72	71	0.12

Appendix Table 62: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From PROMIS Satisfaction w/ Participation in Discretionary Social Activities v1.0 to PROMIS Satisfaction w/ Social Roles & Activities v2.0 (PROsetta Stone Wave 2 Study)-

Note: Table 60 is recommended.

PROMIS Satisfaction w/ Participation in DSA v1.0 T-score Score	PROMIS Satisfaction w/ Participation in DSA v1.0 Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
26.8	12	21	21	20
30.6	13	26	26	27
32.4	14	29	29	30
33.8	15	31	31	31
34.8	16	32	32	32
35.8	17	33	34	34
36.6	18	35	35	35
37.3	19	36	36	35
38.0	20	37	36	36
38.7	21	37	37	37
39.3	22	37	37	37
39.9	23	38	38	38
40.4	24	39	38	39
41.0	25	40	39	39
41.6	26	40	40	40
42.1	27	41	40	40
42.7	28	41	41	41
43.2	29	41	41	41
43.8	30	42	42	42
44.4	31	42	42	42
44.9	32	43	43	43
45.5	33	43	43	43
46.1	34	44	44	44
46.6	35	44	44	44
47.2	36	45	45	45
47.8	37	46	46	46
48.4	38	46	46	46
49.0	39	47	47	47
49.6	40	48	48	48
50.2	41	48	48	48
50.8	42	49	49	49
51.4	43	50	50	50
52.0	44	50	50	50
52.6	45	51	51	51
53.2	46	52	52	52
53.8	47	52	52	52

54.4	48	53	53	53
55.1	49	54	54	54
55.7	50	55	55	55
56.4	51	56	56	56
57.1	52	56	57	56
57.8	53	58	58	57
58.5	54	59	59	58
59.3	55	60	60	59
60.3	56	61	61	60
61.3	57	62	62	62
62.7	58	64	64	64
64.6	59	68	66	66
68.9	60	71	72	77

Appendix Table 63: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for PROMIS Satisfaction w/ Participation in Social Roles v1.0 and PROMIS Satisfaction w/ Social Roles & Activities v2.0 (PROsetta Stone Wave 2 Study) - RECOMMENDED

PROMIS Satisfaction w/ Participation in Social Roles v1.0 T-score	PROMIS Satisfaction w/ Participation in Social Roles v1.0 Raw Score	PROMIS Satisfaction w/ Social Roles & Activities v2.0 T-score	SE
28.4	15	26.3	2.8
29.9	16	28.0	2.5
31.1	17	29.2	2.2
32.1	18	30.3	2.0
32.9	19	31.2	1.9
33.6	20	32.0	1.8
34.2	21	32.7	1.8
34.8	22	33.4	1.7
35.4	23	34.1	1.7
35.9	24	34.1	1.7
36.4	25	35.3	1.7
36.9	26	35.8	1.7
37.4	27	36.4	1.7
37.9	28	37.0	1.6
38.4	29	37.5	1.6
38.8	30	38.0	1.6
39.3	31	38.5	1.6
39.8	32	39.1	1.6
40.2	33	39.6	1.6
40.7	34	40.1	1.6
41.2	35	40.6	1.6
41.6	36	41.1	1.6
42.1	37	41.6	1.6
42.6	38	42.1	1.6
43.1	39	42.6	1.6
43.6	40	43.1	1.6
44.0	41	43.6	1.6
44.5	42	44.2	1.6
45.0	43	44.7	1.6
45.5	44	45.2	1.6
46.0	45	45.7	1.7
46.6	46	46.2	1.7
47.1	47	46.8	1.7

47.6	48	47.3	1.7
48.2	49	47.8	1.7
48.7	50	48.4	1.7
49.2	51	48.9	1.7
49.8	52	49.5	1.7
50.4	53	50.0	1.7
50.9	54	50.6	1.7
51.5	55	51.1	1.7
52.1	56	51.7	1.7
52.7	57	52.3	1.7
53.3	58	52.9	1.7
53.9	59	53.6	1.7
54.6	60	54.2	1.7
55.2	61	54.9	1.7
55.9	62	55.6	1.8
56.5	63	56.3	1.8
57.3	64	57.1	1.9
58.1	65	58.0	2.0
59.0	66	59.0	2.1
60.0	67	60.2	2.3
61.4	68	61.7	2.7
63.4	69	63.7	3.1
67.9	70	68.2	4.8

Appendix Table 64: Direct (Raw to Scale) Equipercentile Crosswalk Table - From PROMIS Satisfaction w/ Participation in Social Roles v1.0 to PROMIS Satisfaction w/ Social Roles & Activities v2.0 (PROsetta Stone Wave 2 Study)- Note: Table 63 is recommended.

PROMIS Satisfaction w/ Participation in Social Roles v1.0 T-score Score	PROMIS Satisfaction w/ Participation in Social Roles v1.0 Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
24.9	14	20	21	22	0.67
28.4	15	27	26	26	0.67
29.9	16	27	28	28	0.59
31.1	17	29	29	29	1.07
32.1	18	30	30	31	1.11
32.9	19	32	32	32	1.37
33.6	20	33	33	33	0.73
34.2	21	34	34	34	0.67
34.8	22	35	35	35	1.22
35.4	23	36	35	35	0.4
35.9	24	36	36	36	0.38
36.4	25	36	36	36	0.38
36.9	26	37	37	37	0.39
37.4	27	37	37	37	0.41
37.9	28	38	38	38	0.24
38.4	29	38	38	38	0.24
38.8	30	39	39	39	0.38
39.3	31	39	39	39	0.38
39.8	32	40	40	40	0.37
40.2	33	40	40	40	0.35
40.7	34	41	40	40	0.22
41.2	35	41	41	41	0.21
41.6	36	41	41	41	0.22
42.1	37	42	42	42	0.23
42.6	38	42	42	42	0.23
43.1	39	42	43	43	0.23
43.6	40	43	43	43	0.29
44.0	41	44	44	44	0.13
44.5	42	44	44	44	0.12
45.0	43	45	45	45	0.18
45.5	44	45	45	45	0.17
46.0	45	46	46	46	0.24
46.6	46	46	46	46	0.23

47.1	47	46	46	46	0.23
47.6	48	47	47	47	0.29
48.2	49	47	47	47	0.28
48.7	50	48	48	48	0.24
49.2	51	48	48	48	0.22
49.8	52	49	49	49	0.32
50.4	53	49	49	50	0.32
50.9	54	50	50	50	0.25
51.5	55	51	51	51	0.31
52.1	56	52	52	51	0.15
52.7	57	52	52	52	0.12
53.3	58	52	53	53	0.12
53.9	59	53	53	53	0.31
54.6	60	53	54	54	0.3
55.2	61	54	54	54	0.24
55.9	62	54	54	55	0.23
56.5	63	55	55	55	0.54
57.3	64	56	56	56	0.42
58.1	65	56	56	56	0.41
59.0	66	57	57	58	0.67
60.0	67	59	58	59	0.68
61.4	68	60	60	61	0.53
63.4	69	62	63	64	1.45
67.9	70	71	71	71	0.1

Appendix Table 65: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From PROMIS Satisfaction w/ Participation in Social Roles v1.0 to PROMIS Satisfaction w/ Social Roles & Activities v2.0 (PROsetta Stone Wave 2 Study) - Note: Table 63 is recommended.

PROMIS Satisfaction w/ Participation in Social Roles v1.0 T-score Score	PROMIS Satisfaction w/ Participation in Social Roles v1.0 Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
24.9	14	21	22	21
28.4	15	26	26	26
29.9	16	27	28	28
31.1	17	29	29	30
32.1	18	30	31	31
32.9	19	31	32	32
33.6	20	33	33	33
34.2	21	34	34	34
34.8	22	35	35	34
35.4	23	36	35	35
35.9	24	36	36	36
36.4	25	36	36	36
36.9	26	37	37	37
37.4	27	37	37	37
37.9	28	38	38	38
38.4	29	38	38	38
38.8	30	39	38	38
39.3	31	39	39	39
39.8	32	40	40	39
40.2	33	40	40	40
40.7	34	40	40	40
41.2	35	41	41	41
41.6	36	41	41	41
42.1	37	42	42	42
42.6	38	42	42	42
43.1	39	42	43	43
43.6	40	43	43	43
44.0	41	44	44	44
44.5	42	44	44	44
45.0	43	45	45	45
45.5	44	45	45	45
46.0	45	46	46	46
46.6	46	46	46	46
47.1	47	46	46	46

47.6	48	47	47	47
48.2	49	47	47	47
48.7	50	48	48	48
49.2	51	48	48	48
49.8	52	49	49	49
50.4	53	49	49	50
50.9	54	50	50	50
51.5	55	51	51	51
52.1	56	52	52	51
52.7	57	52	52	52
53.3	58	52	52	52
53.9	59	53	53	53
54.6	60	53	53	54
55.2	61	54	54	54
55.9	62	54	54	55
56.5	63	55	55	55
57.3	64	56	56	56
58.1	65	56	56	57
59.0	66	57	57	58
60.0	67	59	58	59
61.4	68	60	60	60
63.4	69	61	62	62
67.9	70	70	69	68

Appendix Table 66: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for NIHToolbox Life Satisfaction and Neuro-QOL Positive Affect & Well-being (PROsetta Stone Wave 2 Study) - RECOMMENDED

NIH Toolbox Life Satisfaction T-Score	NIH Toolbox Life Satisfaction Raw Score	Neuro-QOL Positive Affect & Well-being T-score	SE
19.1	10	24.8	4.4
22.1	11	28.1	3.7
24.1	12	30.2	3.3
25.7	13	31.8	3.2
27.1	14	33.1	3.0
28.3	15	34.3	2.9
29.5	16	35.4	2.8
30.5	17	36.4	2.8
31.5	18	37.3	2.7
32.4	19	38.2	2.7
33.3	20	39.0	2.6
34.2	21	39.8	2.6
35.0	22	40.5	2.6
35.8	23	41.3	2.6
36.6	24	42.0	2.6
37.3	25	42.7	2.6
38.0	26	43.4	2.6
38.7	27	44.1	2.6
39.4	28	44.7	2.6
40.1	29	45.4	2.6
40.8	30	46.1	2.6
41.5	31	46.7	2.6
42.2	32	47.4	2.6
42.8	33	48.0	2.6
43.5	34	48.7	2.6
44.2	35	49.3	2.6
44.9	36	50.0	2.6
45.6	37	50.6	2.6
46.4	38	51.3	2.6
47.1	39	52.0	2.6
47.9	40	52.7	2.7
48.7	41	53.4	2.7
49.5	42	54.1	2.7
50.3	43	54.8	2.7

51.2	44	55.5	2.7
52.0	45	56.3	2.7
53.0	46	57.1	2.8
53.9	47	57.9	2.8
54.9	48	58.7	2.8
55.9	49	59.6	2.8
57.1	50	60.5	2.9
58.2	51	61.4	2.9
59.4	52	62.4	2.9
60.7	53	63.5	3.0
62.0	54	64.6	3.0
63.3	55	65.9	3.1
64.8	56	67.2	3.2
66.4	57	68.7	3.3
68.4	58	70.5	3.5
71.0	59	72.9	3.8
74.6	60	76.2	4.5

Appendix Table 67: Direct (Raw to Scale) Equipercentile Crosswalk Table – From NIHToolbox Life Satisfaction to Neuro-QOL Positive Affect & Well-being – Note: Table 66 is recommended.

NIH Toolbox Life Satisfaction T-Score Score	NIH Toolbox Life Satisfaction Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
19.1	10	26	29	31	2.81
22.1	11	33	33	34	1.56
24.1	12	35	35	35	1.27
25.7	13	36	36	36	0.66
27.1	14	38	37	37	0.30
28.3	15	38	38	38	0.29
29.5	16	38	38	38	0.30
30.5	17	38	39	39	0.32
31.5	18	39	39	39	0.50
32.4	19	40	40	40	0.70
33.3	20	41	41	41	0.45
34.2	21	41	41	41	0.47
35.0	22	42	42	42	0.40
35.8	23	43	43	43	0.48
36.6	24	44	44	43	0.62
37.3	25	44	44	44	0.61
38.0	26	45	45	45	0.35
38.7	27	45	45	45	0.36
39.4	28	46	46	46	0.18
40.1	29	46	46	46	0.18
40.8	30	46	46	46	0.19
41.5	31	47	47	47	0.30
42.2	32	47	47	47	0.30
42.8	33	48	48	48	0.28
43.5	34	48	48	48	0.26
44.2	35	49	49	49	0.42
44.9	36	50	50	50	0.35
45.6	37	51	51	51	0.51
46.4	38	52	52	51	0.30
47.1	39	52	52	52	0.31
47.9	40	53	53	53	0.28
48.7	41	54	53	53	0.21

49.5	42	54	54	54	0.20
50.3	43	54	54	54	0.20
51.2	44	55	55	55	0.25
52.0	45	55	55	55	0.24
53.0	46	56	56	56	0.29
53.9	47	57	57	57	0.42
54.9	48	58	58	58	0.28
55.9	49	58	58	58	0.26
57.1	50	59	59	59	0.46
58.2	51	61	60	61	0.46
59.4	52	61	62	62	0.43
60.7	53	63	63	63	0.69
62	54	64	65	65	0.89
63.3	55	67	66	67	0.72
64.8	56	68	68	69	0.5
66.4	57	72	70	70	0.17
68.4	58	72	71	71	0.14
71	59	72	72	72	0.12
74.6	60	72	77	78	0.1

Appendix Table 68: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From NIH Toolbox Life Satisfaction to Neuro-QOL Positive Affect & Well-being – Note: Table 66 is recommended.

NIH Toolbox Life Satisfaction T-Score	NIH Toolbox Life Satisfaction Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
19.1	10	27	30	31
22.1	11	33	33	34
24.1	12	35	35	35
25.7	13	36	36	36
27.1	14	37	37	37
28.3	15	38	38	38
29.5	16	38	38	38
30.5	17	38	39	39
31.5	18	39	39	40
32.4	19	40	40	40
33.3	20	41	41	41
34.2	21	41	41	42
35.0	22	42	42	42
35.8	23	43	43	43
36.6	24	44	44	44
37.3	25	44	44	44
38.0	26	45	45	45
38.7	27	45	45	45
39.4	28	46	46	46
40.1	29	46	46	46
40.8	30	46	46	46
41.5	31	47	47	47
42.2	32	47	47	47
42.8	33	47	48	48
43.5	34	48	48	48
44.2	35	49	49	49
44.9	36	50	50	50
45.6	37	51	51	51
46.4	38	52	52	52
47.1	39	52	52	52
47.9	40	53	53	53
48.7	41	54	54	53
49.5	42	54	54	54

50.3	43	54	54	54
51.2	44	55	55	55
52.0	45	55	55	55
53.0	46	56	56	56
53.9	47	57	57	57
54.9	48	58	58	58
55.9	49	58	58	58
57.1	50	59	60	60
58.2	51	61	61	60
59.4	52	62	62	62
60.7	53	63	63	63
62.0	54	65	64	64
63.3	55	66	66	65
64.8	56	68	67	67
66.4	57	70	69	68
68.4	58	71	70	71
71.0	59	72	72	75
74.6	60	74	75	80

Appendix Table 69: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for NIHToolbox Meaning and Neuro-QOL Positive Affect & Well-being (PROsetta Stone Wave 2 Study) - RECOMMENDED

NIH Toolbox Meaning T-Score	NIH Toolbox Meaning Raw Score	Neuro-QOL Positive Affect & Well-being T-Score	SE
11.9	18	20.4	4
12.2	19	23	3.5
12.7	20	24.8	3.2
13.3	21	26.2	3
14.1	22	27.4	2.8
14.9	23	28.4	2.7
15.9	24	29.3	2.5
16.9	25	30.2	2.4
17.8	26	31	2.4
18.8	27	31.7	2.3
19.7	28	32.4	2.2
20.5	29	33	2.2
21.3	30	33.6	2.1
22.1	31	34.2	2.1
22.8	32	34.8	2.1
23.6	33	35.4	2.1
24.3	34	35.9	2
25	35	36.4	2
25.6	36	36.9	2
26.3	37	37.4	2
26.9	38	37.9	2
27.6	39	38.4	2
28.2	40	38.9	2
28.8	41	39.4	2
29.5	42	39.9	2
30.1	43	40.3	2
30.7	44	40.8	2
31.3	45	41.3	2
31.9	46	41.7	2
32.5	47	42.2	2
33.1	48	42.6	2
33.7	49	43.1	2
34.3	50	43.6	2
34.9	51	44	2

35.5	52	44.5	2
36.1	53	44.9	2
36.7	54	45.4	2
37.3	55	45.9	2
37.9	56	46.3	2
38.5	57	46.8	2
39.1	58	47.3	2
39.7	59	47.8	2
40.3	60	48.2	2
40.9	61	48.7	2
41.5	62	49.2	2
42.2	63	49.7	2
42.8	64	50.2	2
43.5	65	50.7	2
44.1	66	51.2	2.1
44.8	67	51.8	2.1
45.5	68	52.3	2.1
46.2	69	52.8	2.1
46.9	70	53.4	2.1
47.6	71	54	2.1
48.4	72	54.5	2.1
49.1	73	55.1	2.1
49.9	74	55.7	2.2
50.7	75	56.3	2.2
51.5	76	57	2.2
52.4	77	57.6	2.2
53.2	78	58.3	2.2
54.1	79	59	2.3
55.1	80	59.8	2.3
56	81	60.6	2.4
57.1	82	61.4	2.5
58.2	83	62.4	2.5
59.4	84	63.4	2.7
60.7	85	64.5	2.8
62.1	86	65.7	3
63.8	87	67.1	3.2
65.8	88	68.8	3.5
68.4	89	71.1	4
71.9	90	74.3	4.7

Appendix Table 70: Direct (Raw to Scale) Equipercentile Crosswalk Table – From NIH Toolbox Meaning to Neuro-QOL Positive Affect & Well-being – Note: Table 69 is recommended.

NIH Toolbox Meaning T-Score	NIH Toolbox Meaning Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
11.9	18	23	12	12	0.28
12.2	19	23	17	17	0.28
12.7	20	23	22	23	0.38
13.3	21	26	27	28	1.54
14.1	22	29	28	29	1.7
14.9	23	30	29	29	1.87
15.9	24	30	30	30	0.94
16.9	25	30	31	30	0.94
17.8	26	32	31	31	0.94
18.8	27	32	32	32	1
19.7	28	32	33	32	1
20.5	29	33	33	33	1.52
21.3	30	34	34	34	2.26
22.1	31	35	34	34	1.17
22.8	32	36	35	35	0.64
23.6	33	36	36	35	0.65
24.3	34	36	36	36	0.66
25	35	36	37	36	0.67
25.6	36	38	37	37	0.29
26.3	37	38	38	37	0.29
26.9	38	38	38	38	0.3
27.6	39	38	38	38	0.34
28.2	40	39	39	39	0.51
28.8	41	39	39	39	0.51
29.5	42	40	40	40	0.68
30.1	43	40	40	40	0.68
30.7	44	41	41	41	0.42
31.3	45	41	41	41	0.43
31.9	46	42	42	42	0.37
32.5	47	42	42	42	0.37
33.1	48	42	43	43	0.37
33.7	49	43	43	44	0.43
34.3	50	44	44	44	0.56
34.9	51	45	45	45	0.31

35.5	52	46	45	45	0.17
36.1	53	46	46	46	0.17
36.7	54	46	46	46	0.17
37.3	55	47	47	47	0.26
37.9	56	47	47	47	0.26
38.5	57	48	48	48	0.22
39.1	58	48	48	48	0.22
39.7	59	48	49	49	0.22
40.3	60	49	49	49	0.35
40.9	61	50	50	50	0.33
41.5	62	50	50	50	0.31
42.2	63	51	51	50	0.49
42.8	64	51	51	51	0.46
43.5	65	52	52	51	0.27
44.1	66	52	52	52	0.26
44.8	67	52	52	52	0.27
45.5	68	53	53	53	0.26
46.2	69	53	53	53	0.26
46.9	70	54	54	54	0.19
47.6	71	54	54	54	0.19
48.4	72	54	54	54	0.19
49.1	73	55	55	55	0.24
49.9	74	55	55	55	0.24
50.7	75	56	56	56	0.28
51.5	76	56	56	56	0.27
52.4	77	57	57	57	0.36
53.2	78	58	57	58	0.26
54.1	79	58	58	58	0.24
55.1	80	58	59	59	0.47
56	81	59	59	60	0.42
57.1	82	60	60	61	0.47
58.2	83	61	61	61	0.46
59.4	84	61	62	63	0.43
60.7	85	63	63	64	0.72
62.1	86	64	64	65	0.93
63.8	87	67	66	66	0.72
65.8	88	68	68	68	0.52
68.4	89	68	70	70	0.45
71.9	90	72	73	72	0.15

Appendix Table 71: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From NIH Toolbox Meaning to Neuro-QOL Positive Affect & Well-being – Note: Table 69 is recommended.

NIH Toolbox Meaning T-Score	NIH Toolbox Meaning Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
11.9	18	21	22	21
12.2	19	23	24	22
12.7	20	24	25	24
13.3	21	26	26	24
14.1	22	28	28	27
14.9	23	30	29	28
15.9	24	31	30	30
16.9	25	32	31	31
17.8	26	32	32	32
18.8	27	32	32	32
19.7	28	32	33	33
20.5	29	33	34	34
21.3	30	34	34	34
22.1	31	35	35	35
22.8	32	36	35	35
23.6	33	36	36	36
24.3	34	36	36	36
25	35	36	37	37
25.6	36	37	37	37
26.3	37	38	38	38
26.9	38	38	38	38
27.6	39	38	38	39
28.2	40	39	39	39
28.8	41	39	39	40
29.5	42	40	40	40
30.1	43	40	40	41
30.7	44	41	41	41
31.3	45	41	41	42
31.9	46	42	42	42
32.5	47	42	42	43
33.1	48	42	43	43
33.7	49	43	44	44
34.3	50	44	44	44
34.9	51	45	45	45
35.5	52	46	45	45
36.1	53	46	46	46

36.7	54	46	46	46
37.3	55	47	47	47
37.9	56	47	47	47
38.5	57	48	48	48
39.1	58	48	48	48
39.7	59	48	49	49
40.3	60	49	49	49
40.9	61	50	50	50
41.5	62	50	50	50
42.2	63	51	51	50
42.8	64	51	51	51
43.5	65	52	52	51
44.1	66	52	52	52
44.8	67	52	52	52
45.5	68	53	53	53
46.2	69	53	53	53
46.9	70	54	54	54
47.6	71	54	54	54
48.4	72	54	54	54
49.1	73	55	55	55
49.9	74	55	55	56
50.7	75	56	56	56
51.5	76	56	56	56
52.4	77	57	57	57
53.2	78	58	58	58
54.1	79	58	58	58
55.1	80	59	59	59
56	81	59	60	59
57.1	82	60	60	60
58.2	83	61	61	61
59.4	84	62	62	62
60.7	85	62	63	62
62.1	86	64	64	63
63.8	87	66	65	64
65.8	88	68	67	66
68.4	89	70	69	68
71.9	90	73	74	74

Appendix Table 72: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for NIH Toolbox Positive Affect and Neuro-QOL Positive Affect & Well-being (PROsetta Stone Wave 2 Study) - RECOMMENDED

NIH Toolbox Positive Affect T-Score	NIH Toolbox Positive Affect Raw Score	Neuro-QOL Positive Affect & Well-being T-score	SE
14.1	20	25	4
15.3	21	28.4	2.8
16.7	22	30.1	2.4
18.0	23	31.4	2.1
19.3	24	32.4	1.9
20.4	25	33.2	1.7
21.4	26	34	1.6
22.4	27	34.6	1.5
23.2	28	35.2	1.5
24.0	29	35.8	1.4
24.8	30	36.3	1.4
25.5	31	36.8	1.4
26.2	32	37.2	1.4
26.9	33	37.7	1.3
27.5	34	38.1	1.3
28.1	35	38.6	1.3
28.7	36	39	1.3
29.3	37	39.4	1.3
29.8	38	39.8	1.3
30.4	39	40.2	1.3
30.9	40	40.6	1.3
31.4	41	41	1.3
32.0	42	41.4	1.3
32.5	43	41.8	1.3
33.0	44	42.1	1.3
33.5	45	42.5	1.3
34.0	46	42.9	1.3
34.5	47	43.3	1.3
35.0	48	43.7	1.3
35.5	49	44	1.3
36.0	50	44.4	1.3
36.5	51	44.8	1.3
37.0	52	45.2	1.3
37.5	53	45.6	1.3
37.9	54	45.9	1.3
38.4	55	46.3	1.3
38.9	56	46.7	1.3
39.4	57	47.1	1.3
39.9	58	47.4	1.3

40.4	59	47.8	1.3
40.9	60	48.2	1.3
41.3	61	48.6	1.3
41.8	62	49	1.3
42.3	63	49.4	1.3
42.8	64	49.7	1.3
43.3	65	50.1	1.3
43.8	66	50.5	1.3
44.3	67	50.9	1.3
44.7	68	51.3	1.3
45.2	69	51.7	1.3
45.7	70	52.1	1.3
46.2	71	52.5	1.3
46.7	72	52.9	1.3
47.2	73	53.3	1.3
47.8	74	53.7	1.3
48.3	75	54.1	1.3
48.8	76	54.5	1.3
49.3	77	54.9	1.3
49.8	78	55.3	1.3
50.4	79	55.7	1.3
50.9	80	56.1	1.3
51.4	81	56.6	1.3
52.0	82	57	1.3
52.6	83	57.4	1.3
53.1	84	57.9	1.3
53.7	85	58.3	1.3
54.3	86	58.8	1.3
54.9	87	59.2	1.3
55.6	88	59.7	1.3
56.2	89	60.2	1.4
56.9	90	60.7	1.4
57.7	91	61.2	1.4
58.4	92	61.8	1.5
59.3	93	62.4	1.5
60.2	94	63	1.6
61.3	95	63.8	1.7
62.5	96	64.6	1.8
63.9	97	65.6	2.1
65.7	98	66.9	2.4
67.9	99	68.8	2.9
71.4	100	72.5	4.2

Appendix Table 73: Direct (Raw to Scale) Equipercentile Crosswalk Table – From NIHToolbox Positive Affect to Neuro-QOL Positive Affect & Well-being – Note: Table 72 is recommended.

NIH Toolbox Positive Affect T-Score	NIH Toolbox Positive Affect Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
14.1	20	24	17	17	2.09
15.3	21	32	31	31	0.86
16.7	22	32	32	32	0.86
18.0	23	32	33	33	0.86
19.3	24	33	33	33	1.25
20.4	25	34	34	34	1.87
21.4	26	35	35	34	1.05
22.4	27	35	35	35	1.05
23.2	28	36	36	36	0.56
24.0	29	36	36	36	0.56
24.8	30	36	36	37	0.54
25.5	31	37	37	37	1.49
26.2	32	38	37	37	0.23
26.9	33	38	38	38	0.24
27.5	34	38	38	38	0.26
28.1	35	38	38	39	0.27
28.7	36	38	39	39	0.29
29.3	37	39	39	39	0.48
29.8	38	39	40	40	0.49
30.4	39	40	40	40	0.63
30.9	40	41	41	41	0.41
31.4	41	41	41	41	0.37
32.0	42	42	42	42	0.31
32.5	43	42	42	42	0.29
33.0	44	42	42	42	0.31
33.5	45	43	43	43	0.39
34.0	46	43	43	43	0.39
34.5	47	44	44	44	0.49
35.0	48	44	44	44	0.49
35.5	49	45	44	44	0.28
36.0	50	45	45	45	0.28
36.5	51	45	45	45	0.29
37.0	52	45	45	45	0.31
37.5	53	46	45	45	0.16

37.9	54	46	46	46	0.15
38.4	55	46	46	46	0.15
38.9	56	46	46	46	0.15
39.4	57	46	46	47	0.16
39.9	58	47	47	47	0.26
40.4	59	47	47	47	0.24
40.9	60	48	48	48	0.23
41.3	61	49	49	48	0.31
41.8	62	50	49	49	0.29
42.3	63	50	50	50	0.27
42.8	64	50	50	50	0.26
43.3	65	51	51	50	0.42
43.8	66	51	51	51	0.38
44.3	67	52	52	51	0.23
44.7	68	52	52	52	0.22
45.2	69	52	52	52	0.22
45.7	70	53	53	52	0.23
46.2	71	53	53	53	0.22
46.7	72	53	53	53	0.22
47.2	73	53	53	53	0.22
47.8	74	54	54	54	0.15
48.3	75	54	54	54	0.15
48.8	76	54	54	54	0.15
49.3	77	55	55	55	0.2
49.8	78	55	55	55	0.19
50.4	79	55	55	56	0.19
50.9	80	56	56	56	0.25
51.4	81	57	56	56	0.33
52.0	82	57	57	57	0.31
52.6	83	58	57	57	0.22
53.1	84	58	58	58	0.21
53.7	85	58	58	58	0.21
54.3	86	59	59	59	0.39
54.9	87	59	59	59	0.36
55.6	88	59	59	59	0.36
56.2	89	60	60	60	0.39
56.9	90	60	60	60	0.36
57.7	91	61	61	61	0.36
58.4	92	61	61	61	0.35
59.3	93	61	61	62	0.34
60.2	94	62	62	62	2.15

61.3	95	63	63	63	0.55
62.5	96	63	63	64	0.53
63.9	97	64	64	65	0.72
65.7	98	65	66	66	0.54
67.9	99	67	68	68	0.54
71.4	100	72	72	71	0.17

Appendix Table 74: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From NIHToolbox Positive Affect to Neuro-QOL Positive Affect & Well-being - Note: Table 72 is recommended.

NIH Toolbox Positive Affect T-Score	NIH Toolbox Positive Affect Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
14.1	20	25	27	27
15.3	21	32	31	31
16.7	22	32	32	32
18.0	23	32	33	33
19.3	24	33	34	34
20.4	25	34	34	34
21.4	26	34	35	35
22.4	27	35	35	35
23.2	28	36	36	36
24.0	29	36	36	36
24.8	30	36	37	36
25.5	31	37	37	37
26.2	32	38	38	37
26.9	33	38	38	38
27.5	34	38	38	38
28.1	35	38	38	38
28.7	36	38	39	39
29.3	37	39	39	39
29.8	38	39	40	40
30.4	39	40	40	40
30.9	40	41	41	41
31.4	41	41	41	41
32.0	42	42	42	42
32.5	43	42	42	42
33.0	44	42	42	42
33.5	45	43	43	43
34.0	46	43	43	43
34.5	47	44	44	44
35.0	48	44	44	44
35.5	49	45	45	44
36.0	50	45	45	45
36.5	51	45	45	45
37.0	52	45	45	45
37.5	53	46	46	45

37.9	54	46	46	46
38.4	55	46	46	46
38.9	56	46	46	46
39.4	57	46	46	46
39.9	58	47	47	47
40.4	59	47	47	47
40.9	60	48	48	48
41.3	61	49	49	48
41.8	62	50	49	49
42.3	63	50	50	50
42.8	64	50	50	50
43.3	65	51	51	50
43.8	66	51	51	51
44.3	67	52	51	51
44.7	68	52	52	52
45.2	69	52	52	52
45.7	70	52	52	52
46.2	71	53	53	53
46.7	72	53	53	53
47.2	73	53	53	53
47.8	74	54	54	54
48.3	75	54	54	54
48.8	76	54	54	54
49.3	77	55	54	55
49.8	78	55	55	55
50.4	79	55	55	55
50.9	80	56	56	56
51.4	81	57	56	56
52.0	82	57	57	57
52.6	83	58	58	57
53.1	84	58	58	58
53.7	85	58	58	58
54.3	86	59	59	59
54.9	87	59	59	59
55.6	88	59	60	60
56.2	89	60	60	60
56.9	90	60	60	60
57.7	91	61	61	61
58.4	92	61	61	61
59.3	93	62	62	62
60.2	94	62	62	62

61.3	95	63	63	63
62.5	96	64	63	64
63.9	97	64	64	64
65.7	98	65	65	66
67.9	99	66	67	67
71.4	100	71	71	71

Appendix Table 75: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QoL Cognitive Function v2.0 to PROMIS Cognitive Function v2.0 (PROsetta Stone Wave 2 Study) – RECOMMENDED

Neuro-QoL Cognitive Function Raw Score	PROMIS / Neuro-QoL T-score	SE
28	13.3	2.2
29	14.3	2.5
30	15.4	2.6
31	16.5	2.7
32	17.6	2.6
33	18.6	2.6
34	19.6	2.5
35	20.4	2.4
36	21.2	2.3
37	21.9	2.2
38	22.6	2.1
39	23.2	2.0
40	23.8	2.0
41	24.3	1.9
42	24.8	1.9
43	25.3	1.8
44	25.8	1.8
45	26.3	1.7
46	26.7	1.7
47	27.1	1.7
48	27.5	1.6
49	27.9	1.6
50	28.3	1.6
51	28.7	1.6
52	29.0	1.5
53	29.4	1.5
54	29.7	1.5
55	30.0	1.5
56	30.4	1.5
57	30.7	1.5
58	31.0	1.4
59	31.3	1.4
60	31.6	1.4
61	31.9	1.4
62	32.2	1.4

63	32.5	1.4
64	32.8	1.4
65	33.1	1.4
66	33.4	1.4
67	33.7	1.4
68	34.0	1.4
69	34.3	1.4
70	34.6	1.4
71	34.9	1.4
72	35.1	1.4
73	35.4	1.4
74	35.7	1.4
75	36.0	1.4
76	36.3	1.4
77	36.6	1.4
78	36.9	1.4
79	37.1	1.4
80	37.4	1.4
81	37.7	1.4
82	38.0	1.4
83	38.3	1.4
84	38.6	1.4
85	38.9	1.4
86	39.1	1.4
87	39.4	1.4
88	39.7	1.4
89	40.0	1.4
90	40.3	1.4
91	40.6	1.4
92	40.9	1.4
93	41.2	1.4
94	41.5	1.4
95	41.8	1.4
96	42.1	1.4
97	42.4	1.4
98	42.7	1.4
99	43.0	1.4
100	43.3	1.4
101	43.6	1.4
102	43.9	1.4
103	44.3	1.4

104	44.6	1.4
105	44.9	1.4
106	45.2	1.5
107	45.5	1.5
108	45.9	1.5
109	46.2	1.5
110	46.5	1.5
111	46.9	1.5
112	47.2	1.5
113	47.6	1.5
114	47.9	1.5
115	48.3	1.5
116	48.6	1.5
117	49.0	1.5
118	49.4	1.5
119	49.8	1.5
120	50.1	1.5
121	50.5	1.6
122	51.0	1.6
123	51.4	1.6
124	51.8	1.6
125	52.3	1.7
126	52.8	1.7
127	53.3	1.7
128	53.8	1.8
129	54.4	1.9
130	55.0	1.9
131	55.6	2.0
132	56.3	2.1
133	57.1	2.3
134	58.0	2.4
135	59.0	2.6
136	60.2	2.9
137	61.6	3.1
138	63.3	3.5
139	65.6	4.0
140	69.5	5.1

Appendix Table 76: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Cognitive Function v2.0 to PROMIS Cognitive Function v2.0 - Note: Table 75 is recommended.

Neuro-QoL Cognitive Function Raw Score	Equipercentile Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
28	15	10	10	0.61
29	15	11	11	0.61
30	15	12	12	0.61
31	15	13	13	0.61
32	15	14	14	0.61
33	15	15	15	0.61
34	16	16	16	2.00
35	17	17	17	2.00
36	18	18	19	2.00
37	18	19	20	2.00
38	20	20	21	2.00
39	24	21	22	2.00
40	24	22	23	2.00
41	24	23	24	1.00
42	25	24	25	1.00
43	25	25	25	1.06
44	26	25	25	1.22
45	26	25	26	0.61
46	26	26	26	0.62
47	26	26	26	0.66
48	26	26	27	0.66
49	26	27	27	0.62
50	26	27	27	0.62
51	28	27	28	1.41
52	28	28	28	1.27
53	28	28	29	1.27
54	29	29	29	0.72
55	29	29	29	0.72
56	29	30	30	0.72
57	29	30	30	0.72
58	30	30	30	2.00
59	30	31	31	1.41
60	32	31	31	0.23
61	32	32	32	0.21

62	32	32	32	0.25
63	32	32	32	0.31
64	32	33	33	0.39
65	33	33	33	0.89
66	34	33	33	0.51
67	34	34	34	0.52
68	34	34	34	0.52
69	35	34	34	0.29
70	35	35	34	0.30
71	35	35	35	0.32
72	35	35	35	0.33
73	35	35	35	0.37
74	36	36	36	0.29
75	36	36	36	0.28
76	36	36	36	0.28
77	37	37	37	0.30
78	37	37	37	0.32
79	37	37	37	0.32
80	38	38	38	0.22
81	38	38	38	0.21
82	38	38	38	0.21
83	39	39	38	0.15
84	39	39	39	0.14
85	39	39	39	0.15
86	40	39	39	0.21
87	40	40	40	0.21
88	40	40	40	0.21
89	40	40	40	0.20
90	40	41	40	0.20
91	41	41	41	0.22
92	41	41	41	0.21
93	41	41	41	0.21
94	42	42	42	0.36
95	42	42	42	0.33
96	42	42	42	0.33
97	43	43	43	0.23
98	43	43	43	0.22
99	43	43	43	0.21
100	43	43	43	0.21
101	44	44	44	0.32
102	44	44	44	0.32

103	44	44	44	0.32
104	45	45	45	0.39
105	45	45	45	0.40
106	46	45	45	0.23
107	46	46	46	0.24
108	46	46	46	0.23
109	46	46	46	0.24
110	47	47	47	0.29
111	47	47	47	0.28
112	47	47	47	0.28
113	48	48	48	0.19
114	48	48	48	0.19
115	48	48	48	0.19
116	49	49	49	0.23
117	49	49	49	0.23
118	49	49	49	0.23
119	50	50	50	0.31
120	50	50	50	0.30
121	51	51	51	0.22
122	51	51	51	0.21
123	51	51	51	0.21
124	52	52	52	0.50
125	52	52	52	0.47
126	53	53	53	0.32
127	54	53	53	0.23
128	54	54	54	0.21
129	54	54	54	0.19
130	55	55	55	0.41
131	55	55	56	0.38
132	56	56	56	0.22
133	56	57	57	0.63
134	58	58	58	0.35
135	59	59	59	0.36
136	60	60	61	0.42
137	62	62	62	0.29
138	63	64	64	0.24
139	65	66	66	0.22
140	69	70	69	0.08

Appendix Table 77: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Cognitive Function v2.0 to PROMIS Cognitive Function v2.0 - Note: Table 75 is recommended.

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

62	32	32	32
63	32	32	32
64	32	33	33
65	33	33	33
66	34	34	34
67	34	34	34
68	34	34	34
69	35	34	34
70	35	35	35
71	35	35	35
72	35	35	35
73	35	36	36
74	36	36	36
75	36	36	36
76	36	36	36
77	37	37	37
78	37	37	37
79	37	37	37
80	38	38	38
81	38	38	38
82	38	38	38
83	39	39	38
84	39	39	39
85	39	39	39
86	40	40	39
87	40	40	40
88	40	40	40
89	40	40	40
90	41	40	40
91	41	41	41
92	41	41	41
93	41	41	41
94	42	42	42
95	42	42	42
96	42	42	42
97	43	42	43
98	43	43	43
99	43	43	43
100	43	43	44
101	43	44	44
102	44	44	44

103	45	44	44
104	45	45	45
105	45	45	45
106	46	45	45
107	46	46	46
108	46	46	46
109	46	46	46
110	47	47	46
111	47	47	47
112	47	47	47
113	48	48	48
114	48	48	48
115	48	48	48
116	49	49	48
117	49	49	49
118	49	49	49
119	50	50	50
120	50	50	50
121	51	50	50
122	51	51	51
123	51	51	51
124	52	52	52
125	52	52	52
126	53	53	52
127	53	53	53
128	54	54	54
129	54	54	54
130	55	55	55
131	55	56	55
132	56	56	56
133	57	57	57
134	58	58	58
135	59	59	59
136	60	60	60
137	62	62	62
138	64	64	64
139	66	66	66
140	70	70	72