PROsetta Stone® Analysis Report

A Rosetta Stone for Patient Reported Outcomes

David Cella , Benjamin D. Schalet, Michael A. 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.

Volume 3 ***September 17, 2016***

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 unit 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 then 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 in order 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 legacy PRO instruments to be expressed as standardized T-score metrics linked to the PROMIS metric. This report (Volume 3) encompasses 8 linking studies based on available pediatric PRO data from NIH Toolbox, Neuro-QoL, and PROsetta Stone Wave 3. 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. Volume 2 included linking studies based on data that were primarily from PROsetta Stone Waves 1 and 2.

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 for administering 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-item to 20-item short forms for each domain, and a 10-item Global Health Scale that includes global ratings of five broad PROMIS domains and general health perceptions. As described in a full issue of *Medical Care* (Cella et al., 2007), the PROMIS items, banks, and short forms were developed using a standardized, rigorous methodology that began with constructing a consensus-based PROMIS domain framework.

All PROMIS banks have been calibrated according to Samejima's (1969) graded response model and are based on large data collections including both general and clinical samples. All PROMIS banks are 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) a small number of full-bank testing cases (approximately 1,000 per bank) from a general population taking one full bank and (b) 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 seven items administered from each bank. The full-bank testing samples were randomly assigned to one of seven different forms. Each form was composed of one or more PROMIS domains (with an exception of Physical Function, where the bank was split over two forms) and one or more legacy measures of the same or related domains.

The PROMIS Wave I data collection design included a number of widely accepted “legacy” measures. The legacy measures used for validation evidence included Buss-Perry Aggression Questionnaire (BPAQ), Center for Epidemiological Studies Depression Scale (CES-D), Mood and Anxiety Symptom Questionnaire (MASQ), Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F), Brief Pain Inventory (BPI), and SF-36. In addition to PROMIS-

[1] www.nihpromis.org

[2] www.nihroadmap.nih.gov

legacy measure pairings for validity assessment (e.g., PROMIS Depression and CES-D), the PROMIS Wave I data allowed for the potential to link over a dozen pairs of measures/subscales. Furthermore, included within each of the PROMIS banks were items from many other existing measures. Depending on the nature and strength of relationship between the measures, various linking procedures can be used to allow for cross-walking of scores. (Note that most of the linking reports based on the PROMIS Wave 1 dataset are included in Volume 1.)

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[3] provides investigators with brief, yet comprehensive measurement tools 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 clinically relevant and psychometrically robust Quality of Life (QOL) assessment tools for adults and children with neurological disorders. The primary goal of this effort, known as Neuro-QoL [4], 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

[3] www.nihtoolbox.org
[4] www.neuroqol.org

Stigma), eight pediatric item banks (Anger, Anxiety, Depression, Fatigue, Pain, Applied Cognition - General Concerns, Social Relations - Interaction with Peers, and Stigma) and two additional pediatric physical function scales (Lower Extremity Function –Mobility, and Upper Extremity Function -Fine Motor, ADL).

3. Legacy Instruments

The following instruments are widely accepted "legacy" measures that have now been linked to PROMIS instruments. Some of these legacy measures were used as part of the initial validation work for PROMIS and NIH Toolbox; otherwise, they were administered as part of this PROsetta Stone project for specific linking purposes. Data were collected on reference measures (e.g., PROMIS Depression) from a minimum of 400 respondents (for stable item parameter estimation), along with responses to at least one other conceptually similar scale or bank to be linked to the reference measure. (See Table 5.1).

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

The Center for Epidemiological Studies Depression Scale (CES-D) for children is a 20-item measure designed to assess depressive symptoms in the general population. Items are rated for the past week using a four-point scale (from "Not at all" to "A lot"). The CES-D has good psychometric properties and has been used in a variety of contexts. Scores range from 0 to 60 and a score of 15 has been suggested as a possible cut-off for significant levels of depressive symptoms (Weissman, Orvaschel, Padian, 1980; Faulstich et al., 1986).

3.2. Short Mood and Feelings Questionnaire (SMFQ)

The Short Mood and Feelings Questionnaire (SMFQ) is a13-item subscale from a longer 33-item questionnaire (the original MFQ). Each item is rated on a 3-point Likert scale: "true", "sometimes true", and "not true" with respect to events of the past two weeks. The SMFQ is a brief, easy to administer, self-report measure of childhood and adolescent depression, designed for rapid evaluation of core depressive symptomology or for use in epidemiological studies. The questions asked on the survey are based on the DSM-III criteria for depression; however, this instrument should be used as an indicator of depressive symptoms and not as a diagnostic tool; therefore it does not indicate whether a child or adolescent has a particular disorder. This instrument has a companion version, (i.e., the parent/caregiver-report version), which consists of items assessing the same depressive symptoms by a proxy.

3.3. Pediatric Functional Assessment of Chronic Illness Therapy – Fatigue (pedsFACIT-Fatigue)

The Pediatric Functional Assessment of Chronic Illness Therapy-Fatigue (pedsFACIT-F) is a 13-item self-report measure of pediatric cancer-related fatigue and is part of the Functional Assessment of Chronic Illness Therapy (FACIT) measurement system. pedsFACIT-F was derived from the 51-item pediatric fatigue item bank (pedsFIB). Each item is rated on a 0 to 4 scale: "None of the time", "A little bit of the time", "Some of the time", “Most of the time”, “All of the time”. It measures fatigue in the past seven days. pedsFACIT-F deals with the issue of different literacy levels as children age, as well as the differing perceptions of the impact of fatigue over time. Because of the unique aspects of assessing cancer-related fatigue among children, simply modifying the item wording to make the language more developmentally appropriate may not be sufficient for providing clinically useful information in monitoring fatigue over time. pedsFACIT-F has stable measurement properties across age, sex, and cancer types. It can be used in clinical research. Scores on the pedsFACIT-F discriminate between patients with and without anemia and among patients with different functional status.(Lai et al, 2007) As with all FACIT questionnaires, a high score is good. Therefore, a score of “0” indicates a severely symptomatic patient, and the highest possible score (varies per scale and subscale) indicates an asymptomatic patient.

3.4. Pediatric Perceived Cognitive Function Item Bank (Ped 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.

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, with the order of measures presented randomized so as to minimize potential order effects. The original purpose of the PROMIS full-bank administration study was to establish initial validity evidence (e.g., validity coefficients), not to establish linking relationships. Thus, initial analyses of the full-bank administration sample revealed several potential score-linking issues: (a) some measures had severely skewed score distributions; (b) the sample size for some administered measures was relatively small. These score-linking issues can be limiting factors when determining an appropriate linking method (e.g., what method options are available or whether linking can even be conducted). Another potential linking 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 so 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, a person who has a T-score of 60 is one standard deviation higher than the general population for the concept being measured. It therefore follows that, for condition symptoms and negatively-framed or oriented concepts like pain, fatigue, and anxiety, a score of 60 is one standard deviation worse than average; while for functional scores and positively-framed or oriented concepts like physical and social function, a score of 60 is one standard deviation better than average.

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

4.1. IRT Linking

One of the objectives of the current linking analyses is to determine whether the non-PROMIS measures can be added to their respective PROMIS item banks 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 in order that the non-PROMIS item parameters may be placed on the same metric as the PROMIS items; that is, the focus is on placing the parameters of non-PROMIS items on the PROMIS metric. Updating the PROMIS item parameters is not desired, because the larger PROsetta-wide linking exercise is built on the stability of these final PROMIS 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 metric 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 fully necessary (e.g., two samples are from the same population having the same or similar ability levels). Thus, conducting IRT linking would be worthwhile, with potential score accuracy benefits gained.

Once the non-PROMIS items are calibrated on the corresponding PROMIS item bank metric, 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 remains suitable for the aggregate set of items). Thus, prior to conducting IRT linking, it is important to assess the dimensionality of the involved measures based on separate and combined PROMIS and non-PROMIS measures. Various dimensionality assessment tools can be used, including 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 the creation of multidimensionality as a result of aggregating items measuring different traits. For two scales that measure distinct but highly related traits, predicting scores on one scale from those of the other has been used frequently. Concordance tables between PROMIS and non-PROMIS measures can be constructed using equipercentile equating (Lord, 1982; Kolen & Brennan, 2004) when there is insufficient empirical evidence that the instruments measure the same construct. An equipercentile method estimates a nonlinear linking relationship using percentile rank distributions of the two linking measures. The equipercentile linking method can be used in conjunction with a presmoothing method such as the loglinear model (Hanson, Zeng, & Colton, 1994). The frequency distributions are first smoothed using the loglinear model and then equipercentile linking is conducted based on the smoothed frequency distributions of the two measures. Smoothing can also be done at the backend on equipercentile equivalents and is called postsmoothing (Brennan, 2004; Kolen & Brennan, 2004). The cubic-spline smoothing algorithm (Reinsch, 1967) is used in the LEGS program employed in PROsetta analyses (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.

In 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 a 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 scores 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., the 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. This 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 several linking studies using secondary data sets. In each case, we have applied all three linking methods described in sections 4.1 and 4.2. Our purpose is to provide the maximum amount of useful information.

However, the suitability of these methods depends upon the meeting of various linking assumptions. These assumptions require that the two instruments to be linked measure the same construct, show a high correlation, and are relatively invariant in subpopulation differences (Dorans, 2007). The degree to which these assumptions are met varies across linking studies. Given that different researchers may interpret these requirements differently, we have taken a liberal approach for inclusion of linkages in this book. Nevertheless, we recommend that researchers diagnostically review the classical psychometrics and CFA results in light of these assumptions prior to any application of the cross-walk charts or legacy parameters to their own data.

Having investigated a large number of possible links between PROMIS measures 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. For pediatric measures, only one planned link (between PROMIS Pediatric Anxiety and SCARED) failed to reach the .70 criterion.

In other cases, we identified two measures apparently suitable for linking but were unable to obtain sufficient data. That is, we typically sought datasets of a sufficient size (i.e., 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 4.1.1 shows instruments pairs we planned to link but were unable to because the required data were unavailable.

Table 4.1.1 Planned Pediatric Instrument Pairs not Linked - Data Not Available

Planned Instrument Linking Pair	**Reason for Not Linking**
PROMIS Pediatric PF-Mobility Bank and CHAQ	Pair of instruments not administered in single sample
PROMIS Pediatric PF-Upper Extremity Bank and CHAQ	Pair of instruments not administered in single sample
PROMIS Pediatric PF-Upper Extremity Bank and Neuro-QoL Pediatric Upper Extremity	Pair of instruments not administered in single sample
PROMIS Pediatric Fatigue and PedsQL - Fatigue	Sample size < 400
PROMIS Pediatric Fatigue and KidScreen - Physical Well-being	Pair of instruments not administered in single sample
PROMIS Pediatric Pain Interference and Neuro-QoL Pediatric Pain	Sample size < 400
PROMIS Pediatric Pain Interference and PedsQL - Pain	Pair of instruments not administered in single sample
PROMIS Pediatric Pain Interference and KidScreen - Physical Well-being	Pair of instruments not administered in single sample
PROMIS Pediatric Anger and Neuro-QoL Pediatric Anger	Pair of instruments not administered in single sample
PROMIS Pediatric Anger and AESC	Sample size < 400

5. Linking Results

Table 5.1 lists the linking analyses included in this report. These analyses have been conducted based on samples from three different studies: Neuro-QoL, 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 Pediatric Anxiety v1.0	Neuro-QoL Pediatric Anxiety v1.0	Neuro-QoL Wave 1
5.2	PROMIS Pediatric Depressive Symptoms v1.0	CES-D Children	NIH Toolbox CV
5.3	PROMIS Pediatric Depressive Symptoms v1.0	Neuro-QoL Pediatric Depression v1.0	Neuro-QoL Wave1
5.4	PROMIS Pediatric Depressive Symptoms v1.0	SMFQ	NIH Toolbox CV
5.5	PROMIS Pediatric Fatigue v1.0	Pediatric FACIT Fatigue v1.0	PROsetta Stone Wave 3
5.6	PROMIS Pediatric Mobility v1.0	Neuro-QoL Pediatric Mobility v1.0	Neuro-QoL Wave 1
5.7	PROMIS Pediatric Peer Relationships v1.0	Neuro-QoL Pediatric Interaction with Peers v1.0	PROsetta Stone Wave 3
Section	**Neuro-QoL Instrument**	**Instrument to Link**	**Study**
5.8	Neuro-QoL Pediatric Cognitive Function v2.0 *	Pediatric PCF	PROsetta Stone Wave 3

** In 2014, the Neuro-QoL Pediatric Applied Cognition -- General Concerns bank was modified to be the v2.0 Pediatric Cognitive Function item bank.*

5.1. PROMIS Pediatric Anxiety and Neuro-QoL Pediatric Anxiety

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of anxiety, namely, the PROMIS Pediatric Anxiety item bank (three selected items) and Neuro-QoL Pediatric Anxiety (all 19 items). Both instruments were scaled so that higher scores represent higher levels of anxiety. We created raw summed scores for each of the measures separately and then for them 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 513 participants (N = 484 for participants with complete responses).

5.1.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 15 for PROMIS Pediatric Anxiety and 95 for Neuro-QoL Pediatric Anxiety. 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 them 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 Pediatric Anxiety and Neuro-QoL Pediatric Anxiety was 0.91. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pediatric Anxiety and Neuro-QoL Pediatric Anxiety was 0.96. The correlations between the combined score and the measures were 0.91 and 1 for PROMIS Pediatric Anxiety and Neuro-QoL Pediatric Anxiety, respectively.

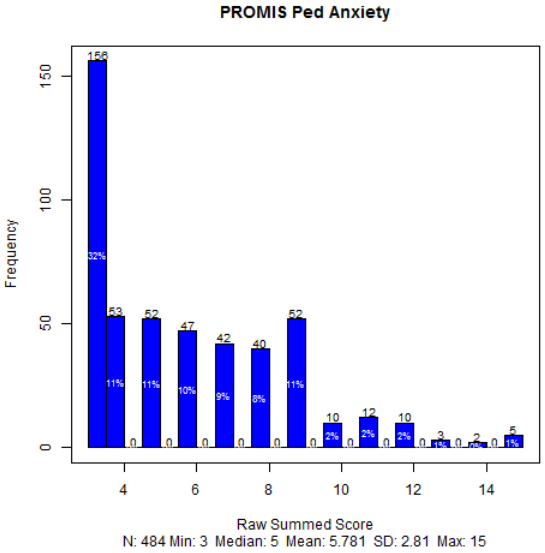


Figure 5.1.1: Raw Summed Score Distribution - PROMIS Pediatric Anxiety

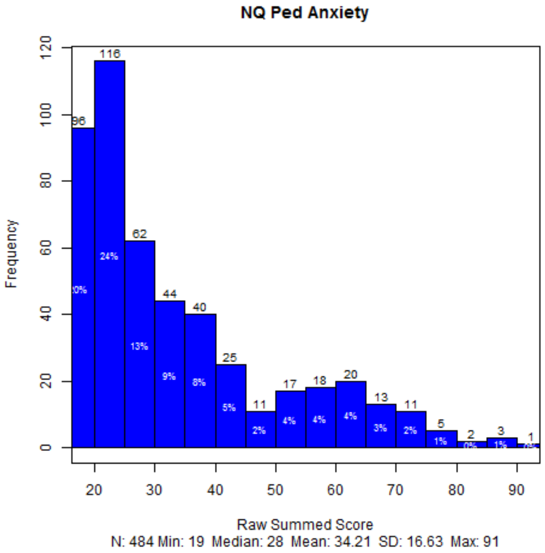


Figure 5.1.2: Raw Summed Score Distribution – Neuro-QoL Pediatric Anxiety

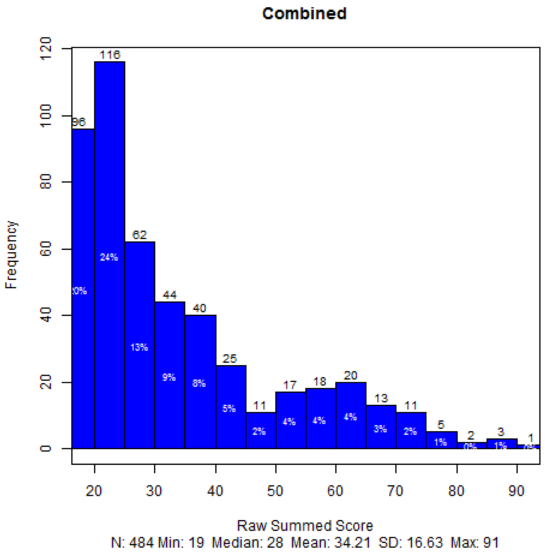


Figure 5.1.3: Raw Summed Score Distribution – Combined

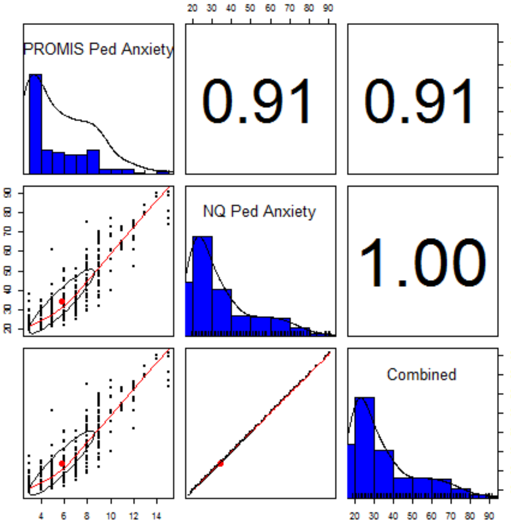


Figure 5.1.4: Scatter Plot Matrix of Raw Summed Scores

5.1.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on them combined. Note that there are three “common items” shared between the two measures (i.e., Neuro-QoL Pediatric Anxiety includes the three items of PROMIS Pediatric Anxiety.) Table 5.1.1 summarizes the results. For PROMIS Pediatric Anxiety, Cronbach’s alpha internal consistency reliability estimate was 0.919 and adjusted (corrected for overlap) item-total correlations ranged from 0.818 to 0.871. For Neuro-QoL Pediatric Anxiety, alpha was 0.972 and adjusted item-total correlations ranged from 0.617 to 0.884. For the 19 total items, alpha was 0.972 and adjusted item-total correlations ranged from 0.617 to 0.884.

Table 5.1.1: Classical Item Analysis

	No.	Alpha	min.r	mean.r	max.r
PROMIS Pediatric Anxiety	3	0.919	0.818	0.839	0.871
Neuro-QoL Pediatric Anxiety	19	0.972	0.617	0.797	0.884
Combined	19	0.972	0.617	0.797	0.884

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 them combined. Table 5.1.2 summarizes the model fit statistics. For PROMIS Pediatric Anxiety, the fit statistics were as follows: CFI = 1, TLI = 1, and RMSEA = 0. For Neuro-QoL Pediatric Anxiety, CFI = 0.982, TLI = 0.98, and RMSEA = 0.107. For the 19 total items, CFI = 0.982, TLI = 0.98, and RMSEA = 0.107. 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 Pediatric Anxiety	3	513	1.000	1.000	0.000
Neuro-QoL Pediatric Anxiety	19	513	0.982	0.980	0.107
Combined	19	513	0.982	0.980	0.107

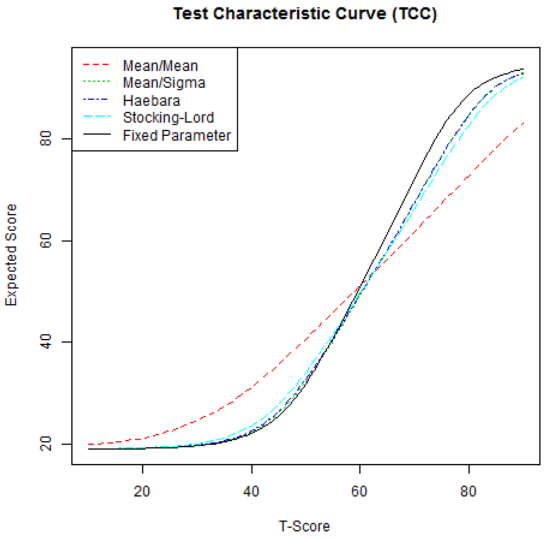
5.1.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 19 items according to the graded response model. The calibration was run using MULTILOG and two different approaches as described previously (i.e., IRT linking vs. fixed-parameter calibration). For IRT linking, all 19 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the three PROMIS Pediatric Anxiety items served as anchor items to transform the item parameter estimates for the Neuro-QoLPediatric Anxiety items onto the PROMIS Pediatric Anxiety metric. We used four IRT linking methods implemented in plink (Weeks, 2010): mean/mean, mean/sigma, Haebara, and Stocking-Lord. The first two methods are based on the mean and standard deviation of item parameter estimates, whereas the latter two are based on the item and test information curves. Table 5.1.3 shows the additive (A) and multiplicative (B) transformation constants derived from the four linking methods. Forfixed-parameter calibration, the item parameters for the PROMIS Pediatric Anxiety items were constrained to their final bank values, while the Neuro-QoL Pediatric Anxiety items were calibrated, under the constraints imposed by the anchor items.

Table 5.1.3: IRT Linking Constants

	A	B
Mean/Mean	2.476	-1.034
Mean/Sigma	1.471	-0.098
Haebara	1.487	-0.122
Stocking-Lord	1.607	-0.233

The item parameter estimates for the Neuro-QoL Pediatric Anxiety items were linked to the PROMIS Pediatric Anxiety metric using the transformation constants shown in Table 5.1.4. The Neuro-QoL Pediatric Anxiety item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Pediatric Anxiety metric. Based on the transformed and fixed-parameter estimates, we derived test characteristic curves (TCC) for Neuro-QoL Pediatric Anxiety 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.



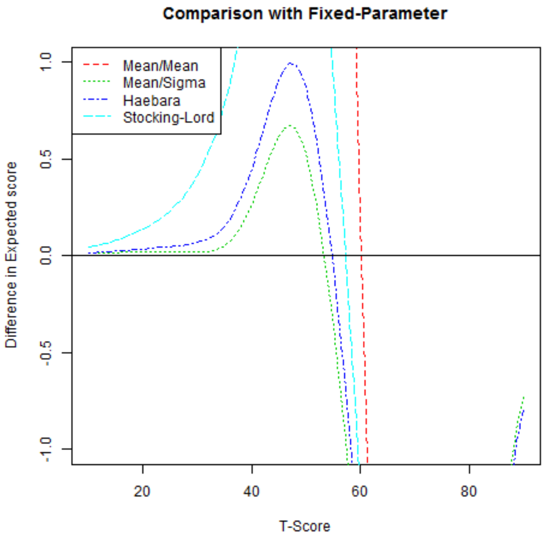


Figure 5.1.7: Test Characteristic Curves (TCC) from Different Linking Methods

Figure 5.1.8: Difference in Test Characteristic Curves (TCC)

Table 5.1.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QoL Pediatric Anxiety. The marginal reliability estimate for Neuro-QoL Pediatric Anxiety based on the item parameter estimates was 0.929. The marginal reliability estimates for PROMIS Pediatric Anxiety and the combined set of items were 0.66 and 0.929, respectively. The slope parameter estimates for Neuro-QoL Pediatric Anxiety ranged from 1.42 to 4.37, with a mean of 2.54. The slope parameter estimates for PROMIS Pediatric Anxiety ranged from 1.51 to 1.81, with a mean of 1.65. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.1.7 displays the scale information functions for PROMIS Pediatric Anxiety, Neuro-QoL Pediatric Anxiety, and the combined set of 19 items. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.1.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.1.4: Fixed-Parameter Calibration Item Parameter Estimates for Neuro-QoL Pediatric Anxiety

a	cb1	cb2	cb3	cb4	NCAT
1.810	-0.784	0.251	1.590	2.650	5
1.640	0.401	1.220	2.610	3.300	5
1.510	-0.853	0.179	1.860	2.850	5
2.572	0.064	0.795	1.848	2.440	5
4.365	0.082	0.697	1.657	2.572	5
3.618	-0.139	0.558	1.637	2.368	5
3.370	-0.204	0.751	1.990	2.493	5
3.430	0.277	1.069	1.751	2.364	5
1.613	-0.946	0.352	1.308	2.336	5
1.419	0.182	1.348	2.168	3.125	5
1.706	-0.069	1.032	1.786	2.757	5
2.082	0.303	0.922	1.545	2.362	5
2.507	0.470	1.039	1.640	2.141	5

2.355	-0.458	0.730	1.540	2.636	5
2.365	-0.371	0.567	1.527	2.420	5
3.024	-0.076	0.866	1.895	2.523	5
2.951	0.429	1.160	2.052	2.676	5
2.855	0.423	1.065	1.705	2.164	5
3.105	0.258	0.979	1.484	2.036	5

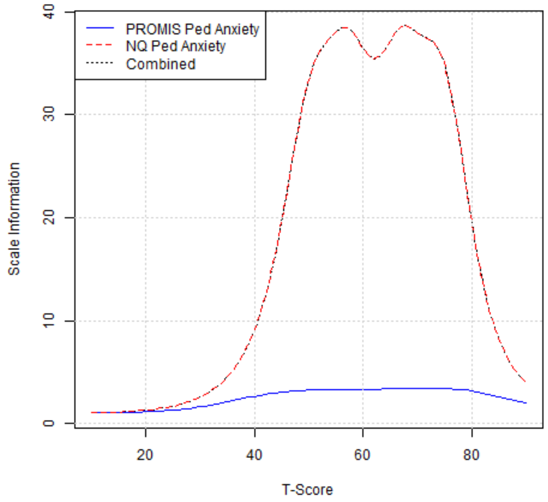


Figure 5.1.9: Comparison of Scale Information Functions

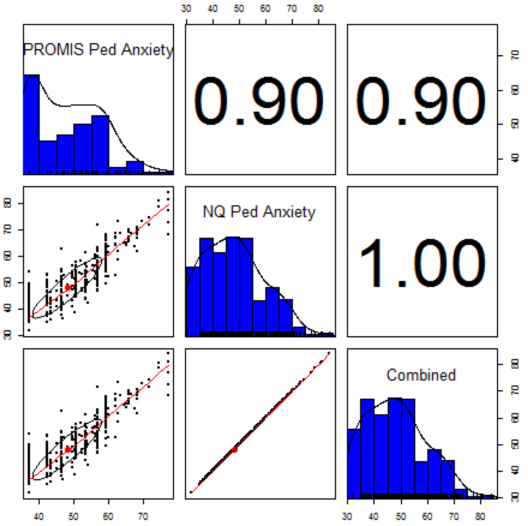


Figure 5.1.10: Comparison of IRT Scaled Scores

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 Neuro-QoL Pediatric Anxiety to a scaled score on PROMIS Pediatric Anxiety can be useful. Based on the Neuro-QoL Pediatric Anxiety 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 Neuro-QoL Pediatric Anxiety to T-score values linked to the PROMIS Pediatric 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 so 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 Neuro-QoL Pediatric Anxiety to a corresponding scaled score on PROMIS Pediatric Anxiety by identifying scores on PROMIS Pediatric Anxiety that have the same percentile ranks as scores on Neuro-QoL Pediatric 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 5.1.9 displays the cumulative distribution functions of the measures. Figure 5.1.10 shows the equipercentile linking functions based on raw summed scores, from Neuro-QoL Pediatric Anxiety to PROMIS Pediatric Anxiety. When the number of raw summed score points differs substantially, the equipercentile linking functions could deviate from each other noticeably. The problem can be exacerbated when the sample size is small. Appendix Table 2 and Appendix Table 3 show the equipercentile crosswalk tables. The result shown in Appendix Table 2 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 3 shows the result based on the indirect (raw summed score to raw summed score equivalent to scaled score equivalent) approach (Refer to Section 4.2 for details). Three separate equipercentile equivalents are presented: one is equipercentile without post smoothing (“Equipercentile Scale Score Equivalents”) and two are 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).

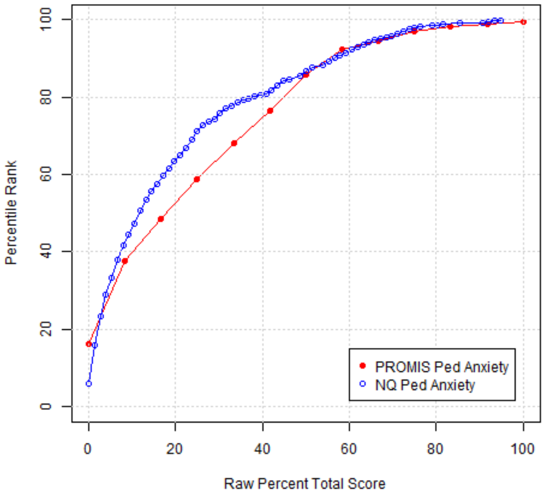


Figure 5.1.11: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

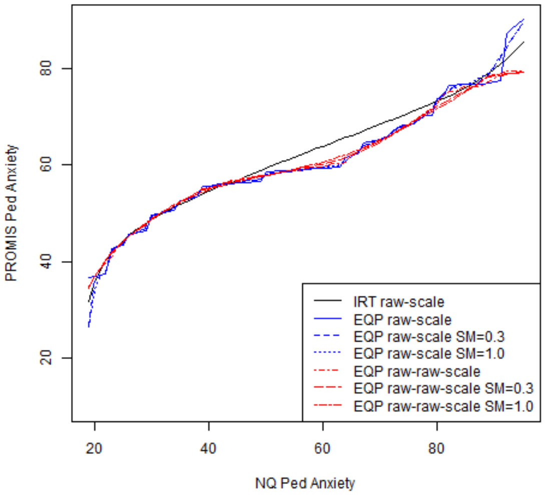


Figure 5.1.12: Equipercentile Linking Functions

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. Another potential issue with IRT-based linking methods is the violation of model assumptions as a result of combining items from two measures (e.g., unidimensionality and local independence). As displayed in Figure 5.1.13, the relationships derived from various linking methods are consistent, which suggests that a robust linking relationship can be determined based on the given sample.

To further facilitate the comparison of the linking methods, Table 5.1.5 reports four statistics summarizing the current sample in terms of the differences between PROMIS Pediatric Anxiety T-scores and Neuro-QoL Pediatric Anxiety scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 5.1.14), 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 (0.909), followed by EQP raw-raw-scale SM=0.0 and EQP raw-raw-scale SM=0.3 (0.908). 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 the smallest RMSD (4.268), followed by EQP raw-raw-scale SM=0.3 (4.28).

Table 5.1.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.899	-0.018	4.970	4.965
IRT raw-scale	0.904	0.052	4.835	4.830
EQP raw-scale SM=0.0	0.905	0.144	4.338	4.336
EQP raw-scale SM=0.3	0.885	1.352	5.398	5.560
EQP raw-scale SM=1.0	0.885	1.547	5.533	5.739
EQP raw-raw-scale SM=0.0	0.908	0.224	4.314	4.315
EQP raw-raw-scale SM=0.3	0.908	0.133	4.283	4.280
EQP raw-raw-scale SM=1.0	0.909	0.129	4.271	4.268

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=484) over a large number of replications (i.e., 10,000).

Table 5.1.6 summarizes the mean and standard deviation of differences between the observed and linked T-scores by linking method and sample size. For each replication, the mean difference between the observed and equated PROMIS Pediatric 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, 0.453. That is, the difference between the mean PROMIS Pediatric Anxiety T-score and the mean equated Neuro-QoL Pediatric Anxiety T-score based on a similar sample of 75 cases is expected to be around ±0.91 (i.e., 2 × 0.453).

Table 5.1.6: Comparison of Resampling Results

Methods	Mean (N=25)	SD (N=25)	Mean (N=50)	SD (N=50)	Mean (N=75)	SD (N=75)
IRT pattern scoring	-0.028	0.978	-0.013	0.663	-0.026	0.528
IRT raw-scale	0.039	0.943	0.051	0.647	0.042	0.515
EQP raw-scale SM=0.0	0.143	0.827	0.128	0.583	0.145	0.463
EQP raw-scale SM=0.3	1.373	1.047	1.359	0.723	1.352	0.571
EQP raw-scale SM=1.0	1.553	1.079	1.545	0.734	1.542	0.585
EQP raw-raw-scale SM=0.0	0.233	0.832	0.218	0.577	0.221	0.460
EQP raw-raw-scale SM=0.3	0.138	0.839	0.128	0.575	0.133	0.459
EQP raw-raw-scale SM=1.0	0.130	0.827	0.129	0.571	0.133	0.453

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 are 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 Pediatric Depressive Symptoms and CES-D Children

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of depression, namely, the PROMIS Pediatric Depressive Symptoms item bank (14 items) and CES-D Children (20 items). Both instruments were scaled so that higher scores represent higher levels of depression. We created raw summed scores for each of the measures separately and then for them 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 N = 1,015 participants.

5.2.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 70 for PROMIS Pediatric Depressive Symptoms and 79 for CES-D Children. 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 them 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 Pediatric Depressive Symptoms and CES-D Children was 0.83. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pediatric Depressive Symptoms and CES-D Children was 0.88. The correlations between the combined score and the measures were 0.96 and 0.96 for PROMIS Pediatric Depressive Symptoms and CES-D Children, respectively.

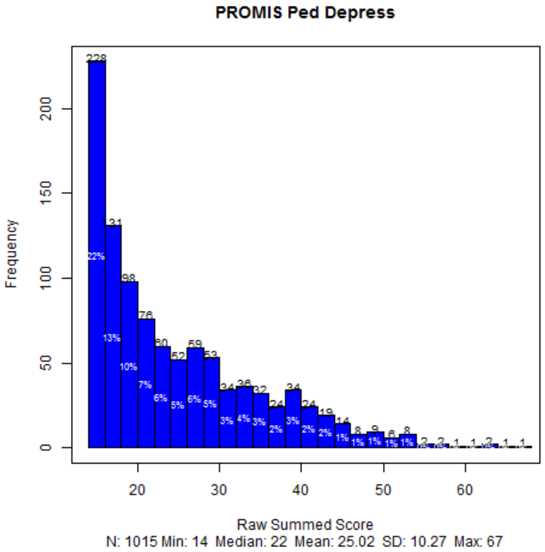


Figure 5.2.1: Raw Summed Score Distribution - PROMIS Pediatric Depressive Symptoms

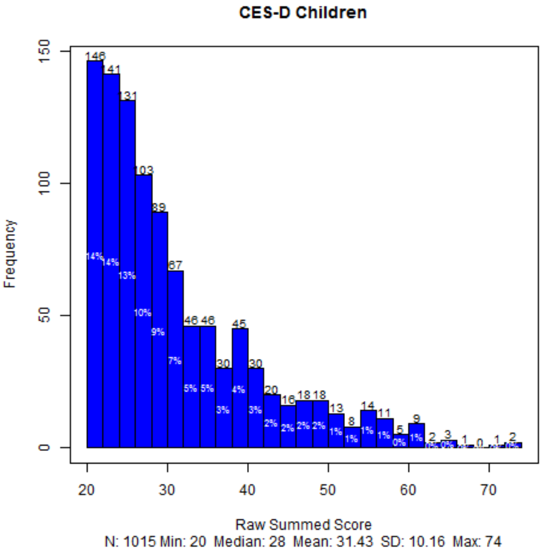


Figure 5.2.2: Raw Summed Score Distribution – CES-D Children

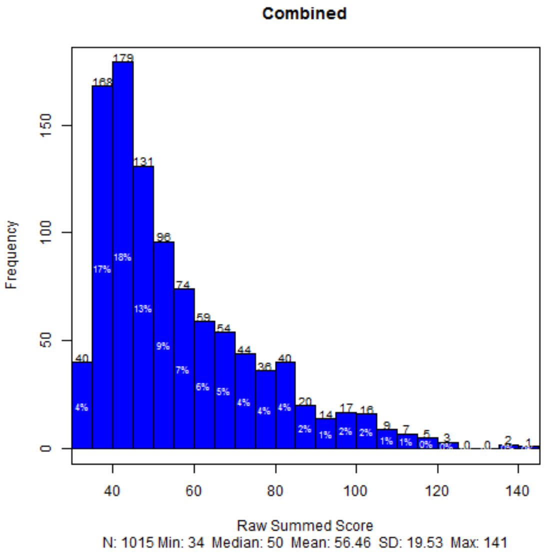


Figure 5.2.3: Raw Summed Score Distribution – Combined

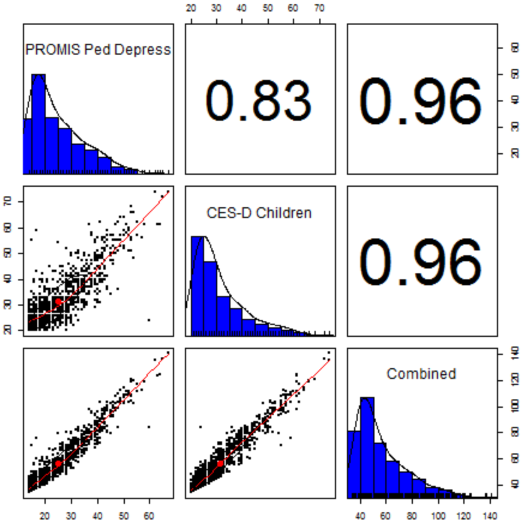


Figure 5.2.4: Scatter Plot Matrix of Raw Summed Scores

5.2.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on them combined. Table 5.2.1 summarizes the results. For PROMIS Pediatric Depressive Symptoms, Cronbach's alpha internal consistency reliability estimate was 0.952 and adjusted (corrected for overlap) item-total correlations ranged from 0.457 to 0.822. For CES-D Children, alpha was 0.928 and adjusted item-total correlations ranged from 0.326 to 0.78. For the 34 items total, alpha was 0.964 and adjusted item-total correlations ranged from 0.304 to 0.803.

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 Pediatric Depressive Symptoms	14	0.952	0.457	0.752	0.822
CES-D Children	20	0.928	0.326	0.613	0.780
Combined	34	0.964	0.304	0.659	0.803

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 them combined. Table 5.2.2 summarizes the model fit statistics. For PROMIS Pediatric Depressive Symptoms, the fit statistics were as follows: CFI = 0.984, TLI = 0.981, and RMSEA = 0.098. For CES-D Children, CFI = 0.94, TLI = 0.933, and RMSEA =

0.098. For the 34 items total, CFI = 0.947, TLI = 0.943, and RMSEA = 0.083. 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 Pediatric Depressive Symptoms	14	1015	0.984	0.981	0.098
CES-D Children	20	1015	0.940	0.933	0.098
Combined	34	1015	0.947	0.943	0.083

5.2.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 34 items according to the graded response model. The calibration was run using MULTILOG and two different approaches as described previously (i.e., IRT linking vs. fixed-parameter calibration). For IRT linking, all 34 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 14 PROMIS Pediatric Depressive Symptoms items served as anchor items to transform the item parameter estimates for the CES-D Children items onto the PROMIS Pediatric Depressive Symptoms 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 Pediatric Depressive Symptoms items were constrained to their final bank values, while the CES-D Children items were calibrated, under the constraints imposed by the anchor items.

Table 5.2.3: IRT Linking Constants

	A	B
Mean/Mean	1.633	-0.681
Mean/Sigma	1.088	-0.024
Haebara	1.067	0.021
Stocking-Lord	1.179	-0.128

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

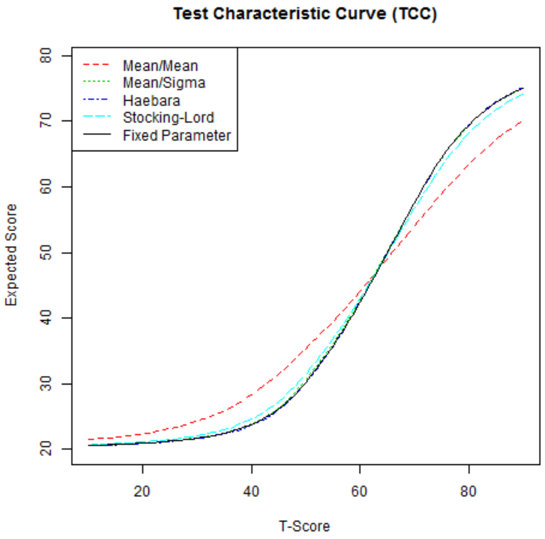


Figure 5.2.5: Test Characteristic Curves (TCC) from Different Linking Methods

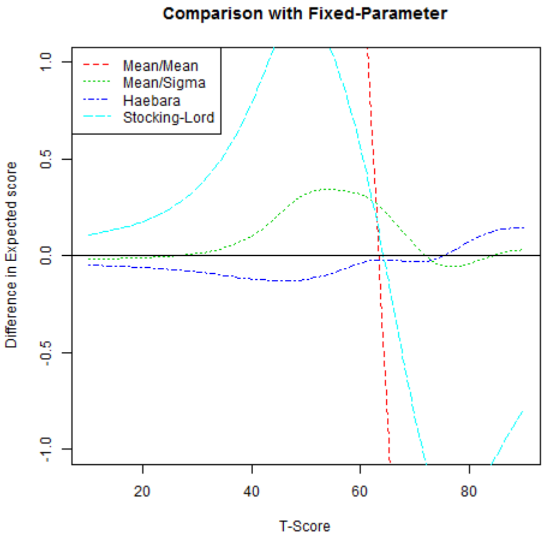


Figure 5.2.6: Difference in Test Characteristic Curves (TCC)

Table 5.2.4 shows the fixed-parameter calibration item parameter estimates for CES-D Children. The marginal reliability estimate for CES-D Children based on the item parameter estimates was 0.893. The marginal reliability estimates for PROMIS Pediatric Depressive Symptoms and the combined set were 0.885 and 0.938, respectively. The slope parameter estimates for CES-D Children ranged from 0.583 to 3.27, with a mean of 1.82. The slope parameter estimates for PROMIS Pediatric Depressive Symptoms ranged from 0.74 to 2.53, with a mean of 1.83. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.2.7 displays the scale information functions for PROMIS Pediatric Depressive Symptoms, CES-D Children, and the combined set of 34 items. 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 CES-D Children

a	cb1	cb2	cb3	NCAT
1.465	0.459	2.001	3.559	4
0.954	0.695	2.473	4.543	4
2.386	0.753	1.710	2.614	4
0.892	0.053	1.681	3.027	4
1.352	-0.175	1.210	2.504	4
3.178	0.024	1.200	2.091	4
1.596	-0.011	1.503	2.714	4
0.583	-3.290	0.439	2.963	4
1.920	0.034	1.438	2.574	4
1.731	0.463	1.801	2.894	4
1.537	0.548	1.861	2.895	4
1.410	0.178	1.829	3.066	4
1.067	0.316	2.001	3.955	4
2.793	0.584	1.398	2.127	4
2.126	0.554	1.492	2.282	4
1.320	0.054	2.040		3
2.374	0.349	1.380	2.355	4
3.274	-0.036	1.243	2.122	4
2.901	0.480	1.372	2.087	4
1.637	-0.122	1.370	2.449	4

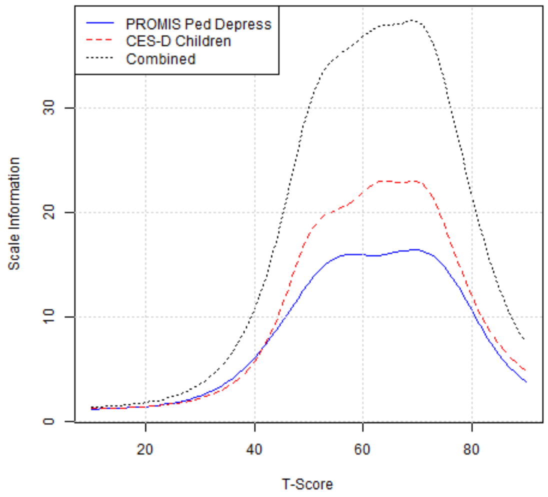


Figure 5.2.7: Comparison of Scale Information Functions

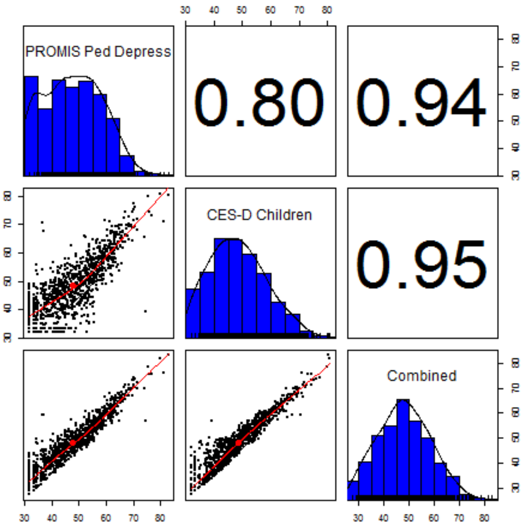


Figure 5.2.8: Comparison of IRT Scaled Scores

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 CES-D Children to a scaled score on PROMIS Pediatric Depressive Symptoms can be useful. Based on the CES-D Children 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 CES-D Children to T-score values linked to the PROMIS Pediatric Depressive Symptoms 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 so 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 CES-D Children to a corresponding scaled score on PROMIS Pediatric Depressive Symptoms by identifying scores on PROMIS Pediatric Depressive Symptoms that have the same percentile ranks as scores on CES-D Children. 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 CES-D Children to PROMIS Pediatric Depressive Symptoms. 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 are 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).

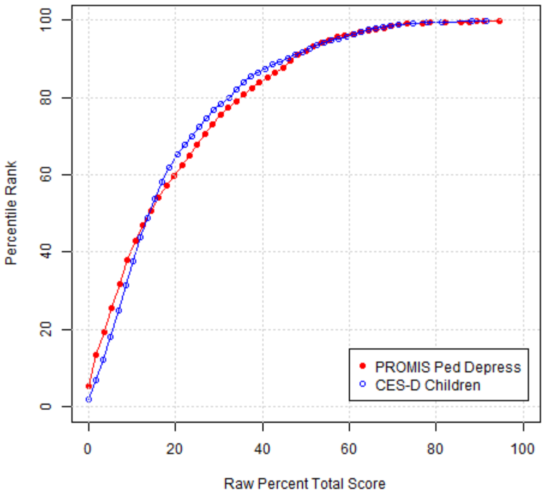


Figure 5.2.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

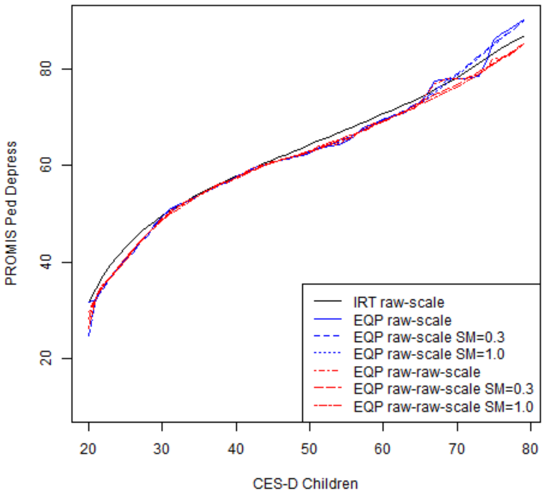


Figure 5.2.10: Equipercentile Linking Functions

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. Another potential issue with IRT-based linking methods is 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 Pediatric Depressive Symptoms T-scores and CES-D Children 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 (0.804), followed by IRT raw-scale, EQP raw-scale SM=0.0, and EQP raw-raw-scale SM=0.0 (0.772). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded the smallest RMSD (6.538), followed by IRT raw-scale (6.995).

Table 5.2.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.804	-0.666	6.507	6.538
IRT raw-scale	0.772	-0.753	6.958	6.995
EQP raw-scale SM=0.0	0.772	0.525	7.092	7.108
EQP raw-scale SM=0.3	0.766	0.798	7.335	7.375
EQP raw-scale SM=1.0	0.766	0.798	7.338	7.378
EQP raw-raw-scale SM=0.0	0.772	0.622	7.124	7.148
EQP raw-raw-scale SM=0.3	0.770	0.707	7.165	7.197
EQP raw-raw-scale SM=1.0	0.768	0.714	7.206	7.238

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 5.2.6 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 Pediatric Depressive Symptoms 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, 0.718. That is, the difference between the mean PROMIS Pediatric Depressive Symptoms T-score and the mean equated CES-D Children T-score based on a similar sample of 75 cases is expected to be around ±1.44 (i.e., 2 × 0.718).

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	-0.669	1.284	-0.662	0.899	-0.673	0.718
IRT raw-scale	-0.743	1.387	-0.760	0.964	-0.744	0.780
EQP raw-scale SM=0.0	0.502	1.403	0.518	0.978	0.526	0.780
EQP raw-scale SM=0.3	0.812	1.437	0.792	1.022	0.789	0.821
EQP raw-scale SM=1.0	0.824	1.459	0.799	1.009	0.803	0.813
EQP raw-raw-scale SM=0.0	0.599	1.414	0.633	0.989	0.620	0.792
EQP raw-raw-scale SM=0.3	0.690	1.410	0.699	0.984	0.712	0.795
EQP raw-raw-scale SM=1.0	0.695	1.436	0.690	0.984	0.713	0.808

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 are sparse. The caveat is that the model should fit the data reasonably well. One of the potential advantages of IRT-based linking is that the item parameters on the linking instrument can be expressed on the metric of the reference instrument and therefore can be combined without significantly altering the underlying trait being measured. As a result, a larger item pool might be available for computerized adaptive testing, or various subsets of items can be used in static short forms. Therefore, IRT-based linking (Appendix Table 4) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.3. PROMIS Pediatric Depressive Symptoms and Neuro-QoL Pediatric Depression

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of depression, namely, the PROMIS Pediatric Depressive Symptoms item bank (eight items) and Neuro-QoL Pediatric Depression (all 17 items). Both instruments were scaled so that higher scores represent higher levels of depression. We created raw summed scores for each of the measures separately and then for them 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 513 participants (N = 494 for participants with complete responses).

5.3.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 40 for PROMIS Pediatric Depressive Symptoms and 85 for Neuro-QoL Pediatric Depression. 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 them 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 Pediatric Depressive Symptoms and Neuro-QoL Pediatric Depression was 0.98. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pediatric Depressive Symptoms and Neuro-QoL Pediatric Depression was 1. The correlations between the combined score and the measures were 0.98 and 1 for PROMIS Pediatric Depressive Symptoms and Neuro-QoL Pediatric Depression, respectively.

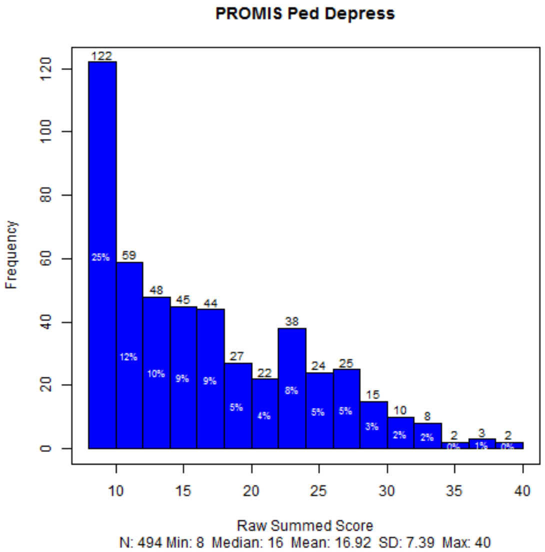


Figure 5.3.1: Raw Summed Score Distribution - PROMIS Pediatric Depressive Symptoms

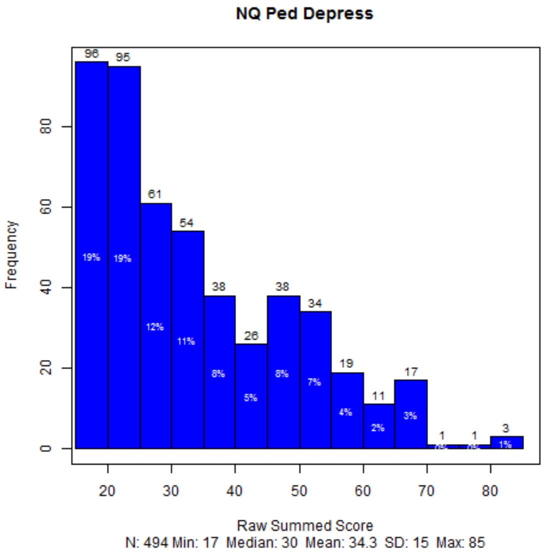


Figure 5.3.2: Raw Summed Score Distribution – Neuro-QoL Pediatric Depression

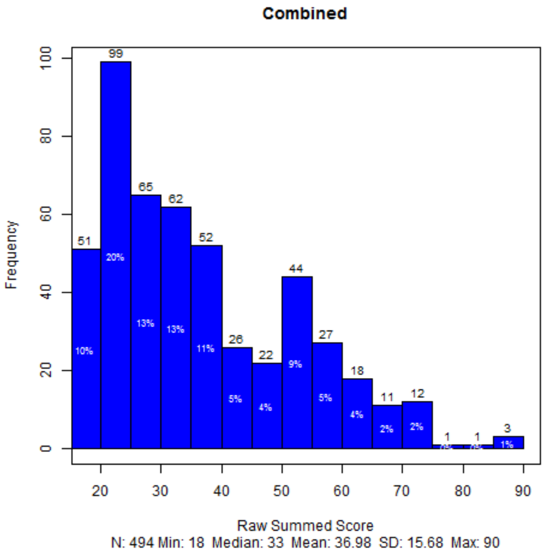


Figure 5.3.3: Raw Summed Score Distribution – Combined

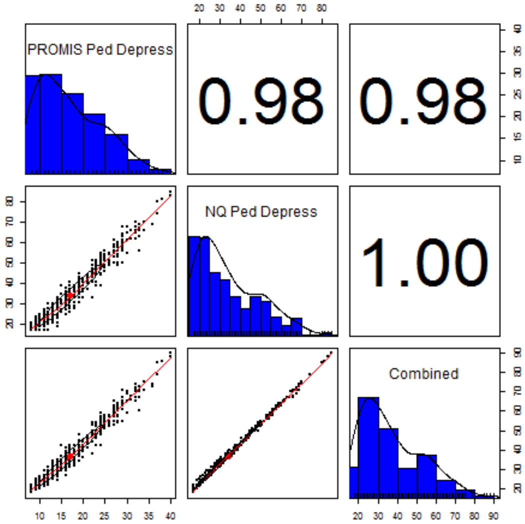


Figure 5.3.4: Scatter Plot Matrix of Raw Summed Scores

5.3.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on them combined. Note that there are seven "common items" shared between the two measures (i.e., Neuro-QoL Pediatric Depression includes seven items from PROMIS Pediatric Depressive Symptoms.) Table 5.3.1 summarizes the results. For PROMIS Pediatric Depressive Symptoms, Cronbach's alpha internal consistency reliability estimate was 0.936 and adjusted (corrected for overlap) item-total correlations ranged from 0.503 to 0.856. For Neuro-QoL Pediatric Depression, alpha was 0.972 and adjusted item-total correlations ranged from 0.611 to 0.885. For the 18 items total, alpha was 0.969 and adjusted item-total correlations ranged from 0.518 to 0.88.

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 Pediatric Depressive Symptoms	8	0.936	0.503	0.781	0.856
Neuro-QoL Pediatric Depression	17	0.972	0.611	0.807	0.885
Combined	18	0.969	0.518	0.790	0.880

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 them combined.

Table 5.3.2 summarizes the model fit statistics. For PROMIS Pediatric Depressive Symptoms, the fit statistics were as follows: CFI = 0.99, TLI = 0.986, and RMSEA = 0.141. For Neuro-QoL Pediatric Depression, CFI = 0.986, TLI = 0.984, and RMSEA = 0.102. For the 18 items total, CFI = 0.985, TLI = 0.983, and RMSEA = 0.097. 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 Pediatric Depressive Symptoms	8	513	0.990	0.986	0.141
Neuro-QoL Pediatric Depression	17	513	0.986	0.984	0.102
Combined	18	513	0.985	0.983	0.097

5.3.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 18 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 18 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the eight PROMIS Pediatric Depressive Symptoms items served as anchor items to transform the item parameter estimates for the Neuro-QoL Pediatric Depression items onto the PROMIS Pediatric Depressive Symptoms 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 Pediatric Depressive Symptoms items were constrained to their final bank values, while the Neuro-QoL Pediatric Depression items were calibrated, under the constraints imposed by the anchor items.

Table 5.3.3: IRT Linking Constants

	A	B
Mean/Mean	1.938	-0.296
Mean/Sigma	1.219	0.242
Haebara	1.173	0.326
Stocking-Lord	1.312	0.206

The item parameter estimates for the Neuro-QoL Pediatric Depression items were linked to the PROMIS Pediatric Depressive Symptoms metric using the transformation constants shown in Table 5.3.3. The Neuro-QoL Pediatric Depression item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Pediatric Depressive Symptoms metric. Based on the transformed and fixed-parameter estimates, we derived test characteristic curves (TCC) for Neuro-QoL Pediatric Depression 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.

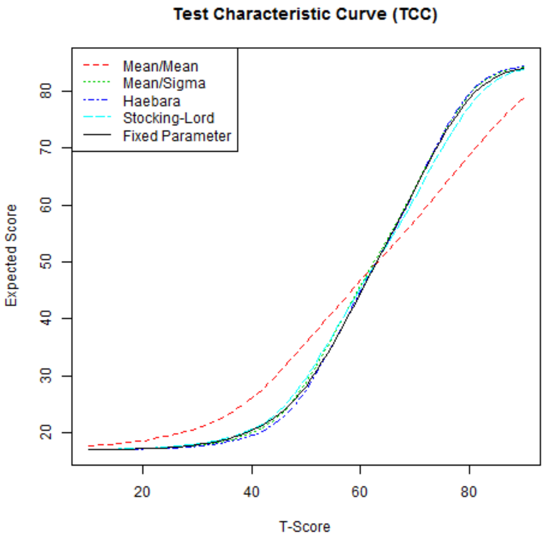


Figure 5.3.5: Test Characteristic Curves (TCC) from Different Linking Methods

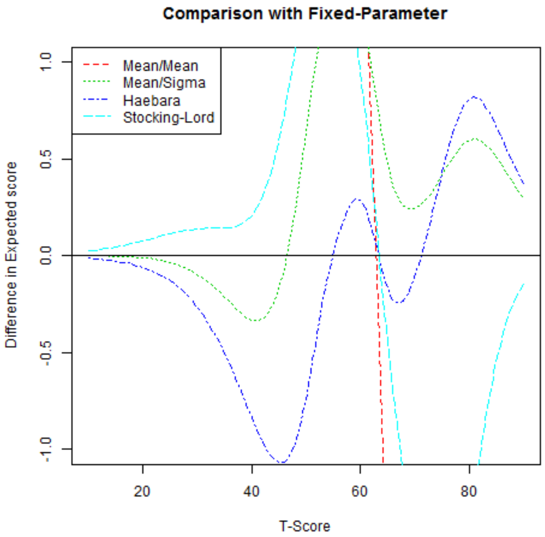


Figure 5.3.6: Difference in Test Characteristic Curves (TCC)

Table 5.3.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QoL Pediatric Depression. The marginal reliability estimate for Neuro-QoL Pediatric Depression based on the item parameter estimates was 0.922. The marginal reliability estimates for PROMIS Pediatric Depressive Symptoms and the combined set were 0.83 and 0.924, respectively. The slope parameter estimates for Neuro-QoL Pediatric Depression ranged from1.56 to 4.23, with a mean of 2.57. The slope parameter estimates for PROMIS Pediatric Depressive Symptoms ranged from 0.739 to 2.46, with a mean of 1.92. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.3.7 displays the scale information functions for PROMIS Pediatric Depressive Symptoms, Neuro-QoL Pediatric Depression, and the combined set of 18 items. 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 Pediatric Depression

a	cb1	cb2	cb3	cb4	NCAT
1.900	-0.747	0.271	1.740	2.750	5
2.040	-0.166	0.629	1.740	2.390	5
1.710	0.306	1.090	2.260	3.000	5
2.420	0.060	0.799	1.700	2.320	5
2.460	0.350	0.959	1.740	2.190	5
2.000	0.253	0.774	1.800	2.410	5

2.110	0.314	0.980	1.910	2.580	5
2.251	0.126	0.926	2.369	3.150	5
3.435	0.128	0.978	2.021	2.710	5
4.061	0.345	1.080	1.893	2.730	5
1.562	-1.658	-0.823	1.096	2.448	5
2.194	-1.052	-0.107	1.440	2.543	5
3.188	-0.079	0.973	1.756	2.562	5
2.063	-0.113	0.877	1.847	2.543	5
3.046	0.401	1.098	2.084	2.672	5
2.940	0.646	1.453	2.233	2.863	5
4.229	0.449	1.119	1.915	2.606	5

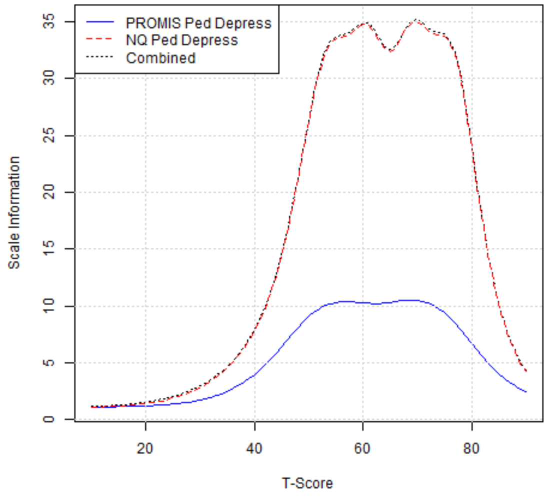


Figure 5.3.7: Comparison of Scale Information Functions

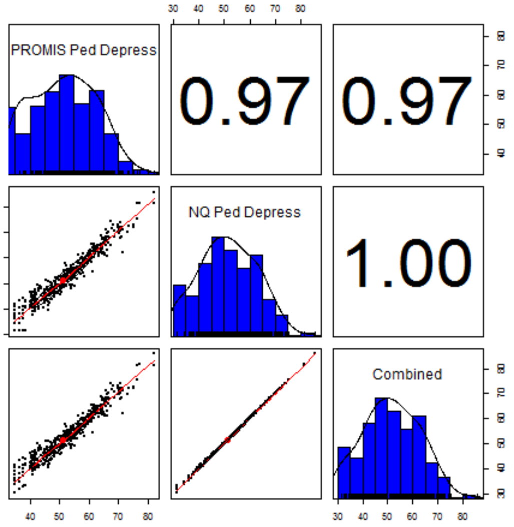


Figure 5.3.8: Comparison of IRT Scaled Scores

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 Pediatric Depression to a scaled score on PROMIS Pediatric Depressive Symptoms can be useful. Based on the Neuro-QoL Pediatric Depression item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 7 can be used to map simple raw summed scores from Neuro-QoL Pediatric Depression to T-score values linked to the PROMIS Pediatric Depressive Symptoms 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 so 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 Pediatric Depression to a corresponding scaled score on PROMIS Pediatric Depressive Symptoms by identifying scores on PROMIS Pediatric Depressive Symptoms that have the same percentile ranks as scores on Neuro-QoL Pediatric Depression. Theoretically, the equipercentile linking function is symmetrical for continuous random variables (X and Y). Therefore, the linking function for the values in X to those in Y is the same as that for the values in Y to those in X. However, for discrete variables like raw summed scores the equipercentile linking functions can be slightly different (due to rounding errors and differences in score ranges) and hence may need to be obtained separately. Figure 5.3.9 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 Pediatric Depression to PROMIS Pediatric Depressive Symptoms. 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 are 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).

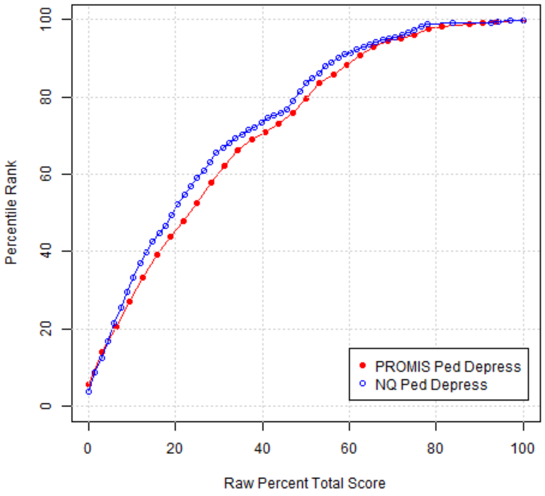


Figure 5.3.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

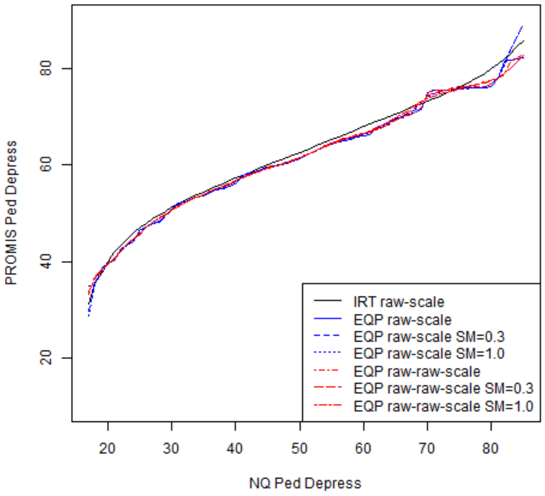


Figure 5.3.10: Equipercentile Linking Functions

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. Another potential issue with IRT-based linking methods is 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 Pediatric Depressive Symptoms T-scores and Neuro-QoL Pediatric Depression 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, EQP raw-scale SM=0.0 produced the best result (0.975), followed by EQP raw-raw-scale SM=0.0, EQP raw-raw-scale SM=0.3, and EQP raw-raw-scale SM=1.0 (0.974). 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 the smallest RMSD (2.482), followed by EQP raw-raw-scale SM=0.3 (2.511).

Table 5.3.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.967	-0.131	2.853	2.853
IRT raw-scale	0.971	-0.012	2.706	2.703
EQP raw-scale SM=0.0	0.975	0.463	2.441	2.482
EQP raw-scale SM=0.3	0.968	0.854	2.854	2.977
EQP raw-scale SM=1.0	0.966	0.937	2.953	3.095
EQP raw-raw-scale SM=0.0	0.974	0.527	2.467	2.520
EQP raw-raw-scale SM=0.3	0.974	0.466	2.470	2.511
EQP raw-raw-scale SM=1.0	0.974	0.490	2.494	2.539

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=494) 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 Pediatric Depressive Symptoms 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, 0.26. That is, the difference between the mean PROMIS Pediatric Depressive Symptoms T-score and the mean equated Neuro-QoL Pediatric Depression T-score based on a similar sample of 75 cases is expected to be around ±0.52 (i.e., 2 × 0.26).

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	-0.136	0.558	-0.131	0.381	-0.134	0.304
IRT raw-scale	-0.009	0.528	-0.014	0.361	-0.014	0.285
EQP raw-scale SM=0.0	0.465	0.475	0.456	0.328	0.460	0.265
EQP raw-scale SM=0.3	0.860	0.552	0.856	0.379	0.856	0.301
EQP raw-scale SM=1.0	0.937	0.580	0.939	0.395	0.942	0.311
EQP raw-raw-scale SM=0.0	0.525	0.477	0.534	0.330	0.523	0.261
EQP raw-raw-scale SM=0.3	0.465	0.481	0.464	0.329	0.466	0.260
EQP raw-raw-scale SM=1.0	0.501	0.488	0.491	0.333	0.492	0.265

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 are 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 Pediatric Depressive Symptoms and SMFQ

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of depression, namely, the PROMIS Pediatric Depressive Symptoms item bank (14 items) and SMFQ (13 items). Both instruments were scaled so that higher scores represent higher levels of depression. We created raw summed scores for each of the measures separately and then for them 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 N = 1,015 participants.

5.4.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 70 for PROMIS Pediatric Depressive Symptoms and 39 for SMFQ. 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 them 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 Pediatric Depressive Symptoms and SMFQ was 0.79. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pediatric Depressive Symptoms and SMFQ was 0.86. The correlations between the combined score and the measures were 0.98 and 0.9 for PROMIS Pediatric Depressive Symptoms and SMFQ, respectively.

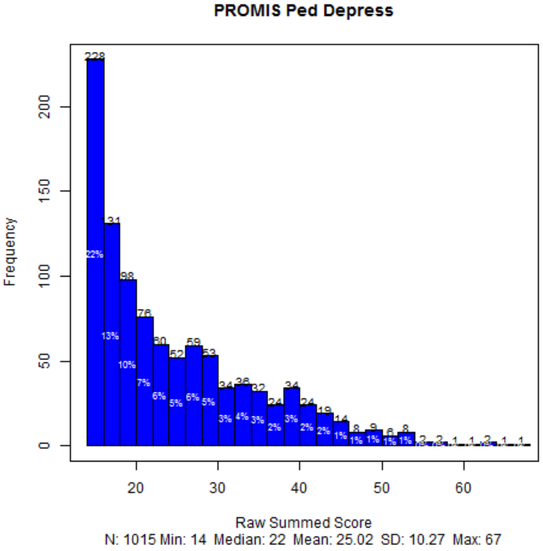


Figure 5.4.1: Raw Summed Score Distribution - PROMIS Pediatric Depressive Symptoms

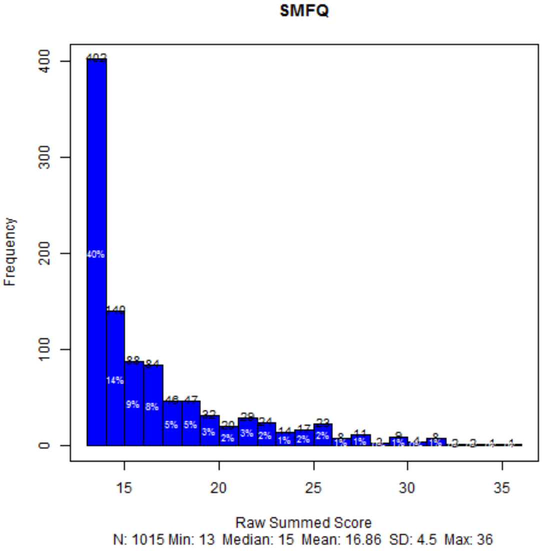


Figure 5.4.2: Raw Summed Score Distribution – SMFQ

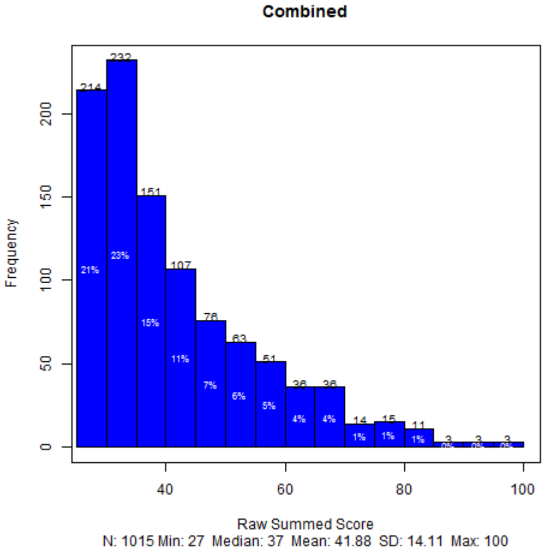


Figure 5.4.3: Raw Summed Score Distribution – Combined

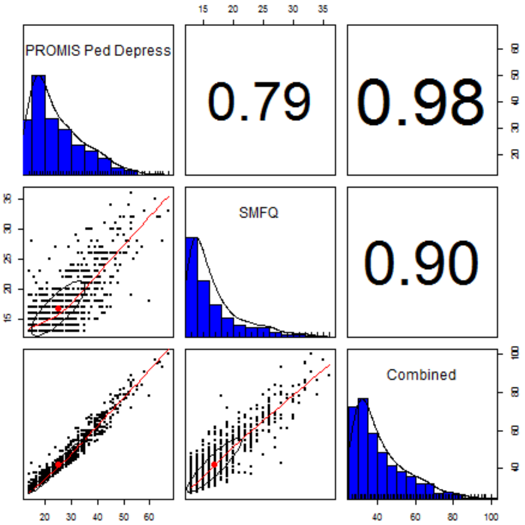


Figure 5.4.4: Scatter Plot Matrix of Raw Summed Scores

5.4.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on them combined. Table 5.4.1 summarizes the results. For PROMIS Pediatric Depressive Symptoms, Cronbach's alpha internal consistency reliability estimate was 0.952 and adjusted (corrected for overlap) item-total correlations ranged from 0.457 to 0.822. For SMFQ, alpha was 0.897 and adjusted item-total correlations ranged from 0.49 to 0.714. For the 27 items total, alpha was 0.956 and adjusted item-total correlations ranged from 0.439 to 0.82.

Table 5.4.1: Classical Item Analysis

			Adjusted (corrected for overlap) Item-total Correlation		
	No. Items	Cronbach's Alpha Internal Consistency Reliability Estimate	Minimum	Mean	Maximum
PROMIS Pediatric Depression	14	0.952	0.457	0.752	0.822
SMFQ	13	0.897	0.490	0.613	0.714
Combined	27	0.956	0.439	0.670	0.820

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 them combined. Table 5.4.2 summarizes the model fit statistics. For PROMIS Pediatric Depressive Symptoms, the fit statistics were as follows: CFI = 0.984, TLI = 0.981, and RMSEA = 0.098. For SMFQ, CFI = 0.982, TLI = 0.978, and RMSEA = 0.065. For the 27 items total, CFI = 0.966, TLI = 0.963, and RMSEA = 0.08. 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 Pediatric Depression	14	1015	0.984	0.981	0.098
SMFQ	13	1015	0.982	0.978	0.065
Combined	27	1015	0.966	0.963	0.080

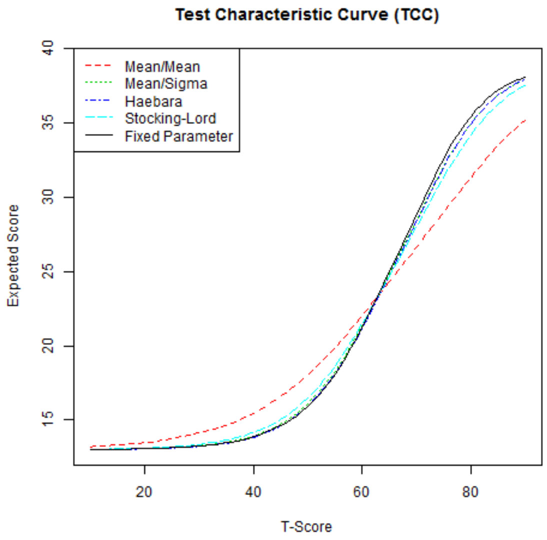
5.4.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 27 items according to the graded response model. The calibration was run using `MULTILOG` and two different approaches as described previously (i.e., IRT linking vs. fixed-parameter calibration). For IRT linking, all 27 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 14 PROMIS Pediatric Depressive Symptoms items served as anchor items to transform the item parameter estimates for the SMFQ items onto the PROMIS Pediatric Depressive Symptoms 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 Pediatric Depressive Symptoms items were constrained to their final bank values, while the SMFQ items were calibrated under the constraints imposed by the anchor items.

Table 5.4.3: IRT Linking Constants

	A	B
Mean/Mean	1.702	-0.807
Mean/Sigma	1.107	-0.074
Haebara	1.091	-0.030
Stocking-Lord	1.207	-0.192

The item parameter estimates for the SMFQ items were linked to the PROMIS Pediatric Depressive Symptoms metric using the transformation constants shown in Table 5.4.3. The SMFQ item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Pediatric Depressive Symptoms metric. Based on the transformed and fixed-parameter estimates, we derived test characteristic curves (TCC) for SMFQ as shown in Figure 5.4.5. Using the fixed-parameter calibration as a basis, we then examined the difference with each of the TCCs from the four linking methods. Figure 5.4.6 displays the differences on the vertical axis.



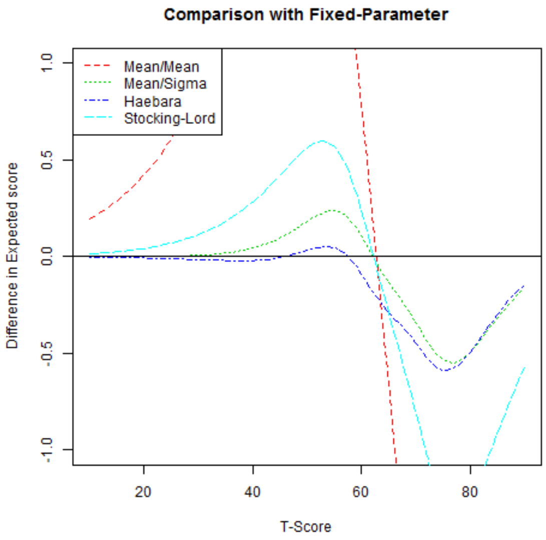


Figure 5.4.5: Test Characteristic Curves (TCC) from Different Linking Methods

Figure 5.4.6: Difference in Test Characteristic Curves (TCC)

Table 5.4.4 shows the fixed-parameter calibration item parameter estimates for SMFQ. The marginal reliability estimate for SMFQ based on the item parameter estimates was 0.785. The marginal reliability estimates for PROMIS Pediatric Depressive Symptoms and the combined set were 0.885 and 0.915, respectively. The slope parameter estimates for SMFQ ranged from 1.1 to 2.96, with a mean of 2.05. The slope parameter estimates for PROMIS Pediatric Depressive Symptoms ranged from 0.74 to 2.53, with a mean of 1.83. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.4.7 displays the scale information functions for PROMIS Pediatric Depressive Symptoms, SMFQ, and the combined set of 27 items. 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 for SMFQ

a	cb1	cb2	NCAT
1.414	0.089	2.249	3
1.702	1.096	2.723	3
1.100	0.474	2.765	3
1.193	0.198	2.647	3
2.886	1.134	2.210	3
1.900	1.357	2.705	3
1.436	0.116	2.152	3
2.954	1.230	2.361	3
2.113	1.438	2.963	3
2.321	0.389	2.164	3
2.964	0.958	2.112	3
2.205	0.593	2.145	3
2.494	0.926	2.478	3

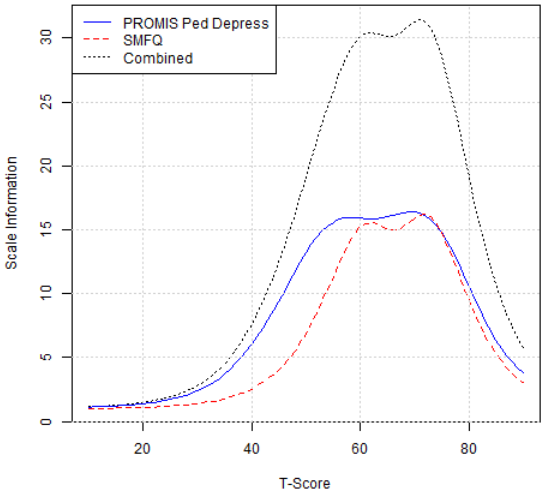


Figure 5.4.7: Comparison of Scale Information Functions

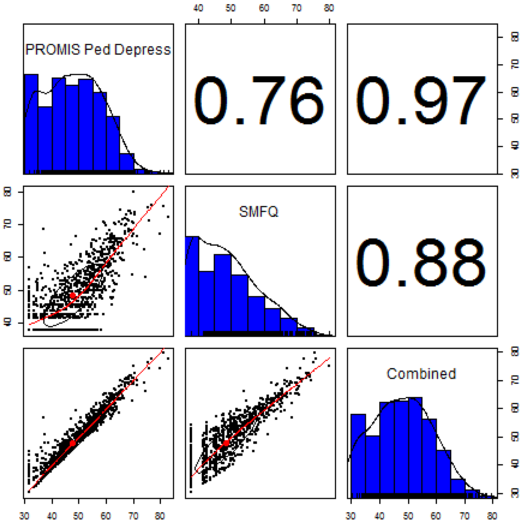


Figure 5.4.8: Comparison of IRT Scaled Scores

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 SMFQ to a scaled score on PROMIS Pediatric Depressive Symptoms can be useful. Based on the SMFQ item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 10 can be used to map simple raw summed scores from SMFQ to T-score values linked to the PROMIS Pediatric Depressive Symptoms 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 so 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 SMFQ to a corresponding scaled score on PROMIS Pediatric Depressive Symptoms by identifying scores on PROMIS Pediatric Depressive Symptoms that have the same percentile ranks as scores on SMFQ. 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 SMFQ to PROMIS Pediatric Depressive Symptoms. When the number of raw summed score points differs substantially, the equipercentile linking functions could deviate from each other noticeably. The problem can be exacerbated when the sample size is small. Appendix Table 11 and Appendix Table 12 show the equipercentile crosswalk tables. The result shown in Appendix Table 11 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 12 shows the result based on the indirect (raw summed score to raw summed score equivalent to scaled score equivalent) approach (Refer to Section 4.2 for details). Three separate equipercentile equivalents are presented: one is equipercentile without post smoothing (“Equipercentile Scale Score Equivalents”) and two are 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).

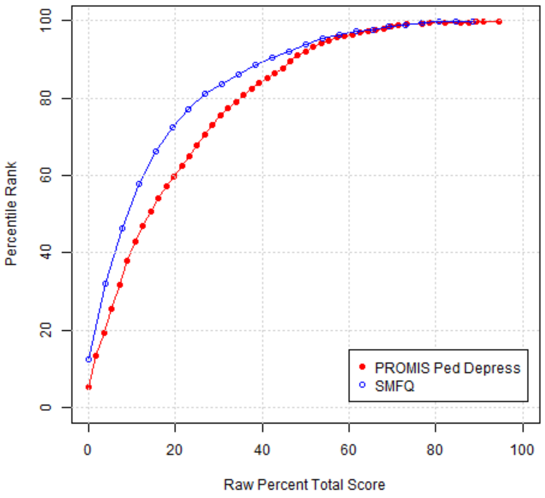


Figure 5.4.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

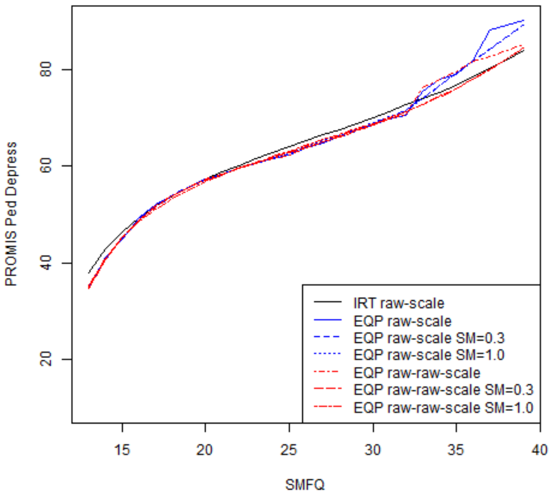


Figure 5.4.10: Equipercentile Linking Functions

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. Another potential issue with IRT-based linking methods is 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 Pediatric Depressive Symptoms T-scores and SMFQ 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 (0.763), followed by IRT raw-scale (0.745). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded the smallest RMSD (7.063), followed by IRT raw-scale (7.304).

Table 5.4.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.763	-0.815	7.019	7.063
IRT raw-scale	0.745	-0.877	7.255	7.304
EQP raw-scale SM=0.0	0.740	0.486	7.531	7.543
EQP raw-scale SM=0.3	0.741	0.488	7.508	7.520
EQP raw-scale SM=1.0	0.742	0.465	7.468	7.479
EQP raw-raw-scale SM=0.0	0.740	0.543	7.531	7.547
EQP raw-raw-scale SM=0.3	0.741	0.589	7.507	7.527
EQP raw-raw-scale SM=1.0	0.743	0.533	7.431	7.446

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 5.4.6 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 Pediatric Depressive Symptoms 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, 0.786. That is, the difference between the mean PROMIS Pediatric Depressive Symptoms T-score and the mean equated SMFQ T-score based on a similar sample of 75 cases is expected to be around ±1.57 (i.e., 2 × 0.786).

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	-0.806	1.370	-0.808	0.965	-0.809	0.786
IRT raw-scale	-0.885	1.431	-0.877	0.998	-0.886	0.817
EQP raw-scale SM=0.0	0.495	1.497	0.497	1.033	0.495	0.827
EQP raw-scale SM=0.3	0.478	1.489	0.497	1.033	0.487	0.830
EQP raw-scale SM=1.0	0.477	1.473	0.453	1.033	0.465	0.832
EQP raw-raw-scale SM=0.0	0.556	1.489	0.545	1.045	0.542	0.834
EQP raw-raw-scale SM=0.3	0.604	1.483	0.590	1.032	0.576	0.836
EQP raw-raw-scale SM=1.0	0.543	1.463	0.526	1.027	0.527	0.817

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 are 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 Pediatric Fatigue and Pediatric FACIT Fatigue

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of fatigue, namely the PROMIS Pediatric Fatigue item bank (15 items) and Peds FACIT Fatigue (13 items). Both instruments were scaled such that higher scores represent higher levels of fatigue. We created raw summed scores for each of the measures separately and then for them 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 507 participants (N = 505 for participants with complete responses).

5.5.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 75 for PROMIS Pediatric Fatigue and 65 for Peds FACIT-F. 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 them 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 Pediatric Fatigue and Peds FACIT-F was 0.86. The disattenuated (corrected for unreliabilies) correlation between PROMIS Pediatric Fatigue and Peds FACIT-F was 0.9. The correlations between the combined score and the measures were 0.97 and 0.95 for PROMIS Pediatric Fatigue and Peds FACIT-F, respectively.

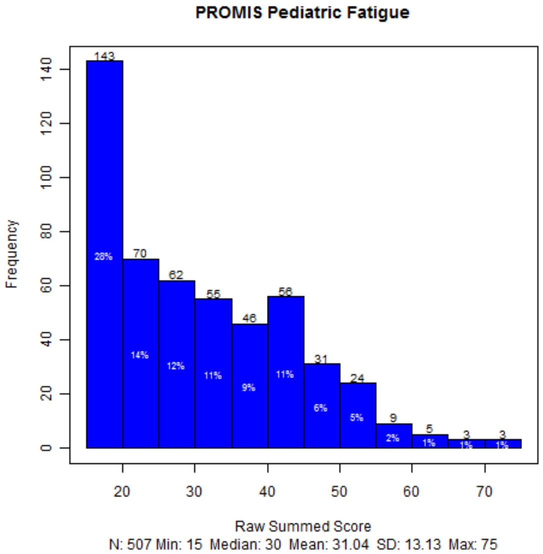


Figure 5.5.1: Raw Summed Score Distribution - PROMIS Peds Fatigue

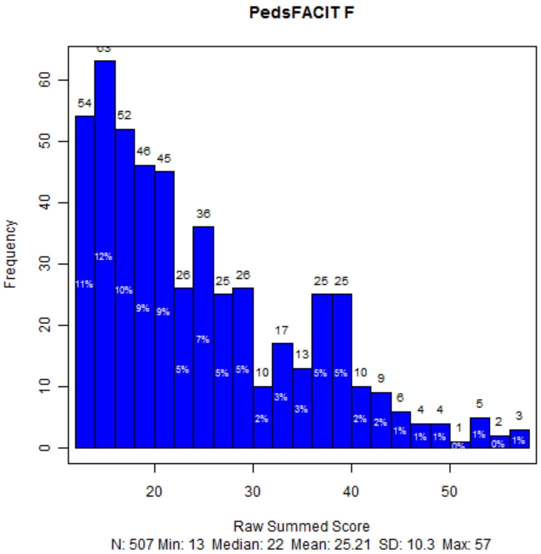


Figure 5.5.2: Raw Summed Score Distribution – Peds FACIT-F

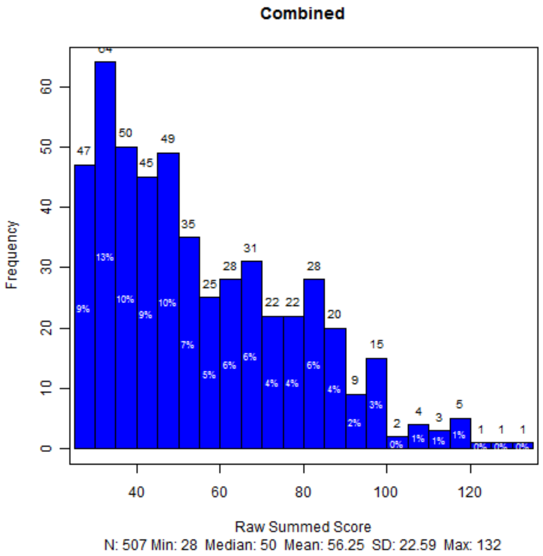


Figure 5.5.3: Raw Summed Score Distribution – Combined

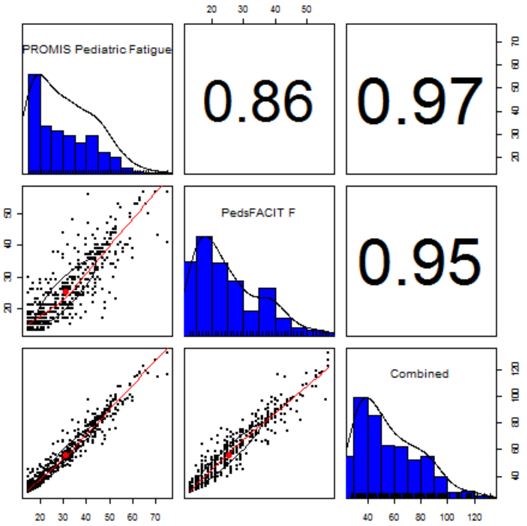


Figure 5.5.4: Scatter Plot Matrix of Raw Summed Scores

5.5.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on them combined. Table 5.5.1 summarizes the results. For PROMIS Pediatric Fatigue, Cronbach's alpha internal consistency reliability estimate was 0.96 and adjusted (corrected for overlap) item-total correlations ranged from 0.576 to 0.833. For Peds FACIT-F, alpha was 0.941 and adjusted item-total correlations ranged from 0.506 to 0.818. For the 28 items, alpha was 0.972 and adjusted item-total correlations ranged from 0.474 to 0.825.

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 Pediatric Fatigue	15	0.960	0.576	0.770	0.833
Peds FACIT-F	13	0.941	0.506	0.718	0.818
Combined	28	0.972	0.474	0.735	0.825

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 Pediatric Fatigue, the fit statistics were as follows: CFI = 0.972, TLI = 0.967, and RMSEA = 0.118. For Peds FACIT-F, CFI = 0.949, TLI = 0.939, and RMSEA = 0.161. For the 28 items, CFI = 0.951, TLI = 0.948, and RMSEA

= 0.102. 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 Pediatric Fatigue	15	507	0.972	0.967	0.118
PedsFACIT F	13	507	0.949	0.939	0.161
Combined	28	507	0.951	0.948	0.102

5.5.4. Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 28 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 28 items were calibrated freely on the conventional theta metric (mean=0, SD=1). Then the 15 PROMIS Pediatric Fatigue items served as anchor items to transform the item parameter estimates for the Peds FACIT-F items onto the PROMIS Pediatric 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 5.5.3 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 Pediatric Fatigue items were constrained to their final bank values, while the Peds FACIT-F items were calibrated, under the constraints imposed by the anchor items.

Table 5.5.3: IRT Linking Constants

	A	B
Mean/Mean	2.173	-0.376
Mean/Sigma	1.453	0.164
Haebara	1.392	0.192
Stocking-Lord	1.618	0.018

The item parameter estimates for the Peds FACIT-F items were linked to the PROMIS Pediatric Fatigue metric using the transformation constants shown in Table 5.5.3. The Peds FACIT-F item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Pediatric Fatigue metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for Peds FACIT-F 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.

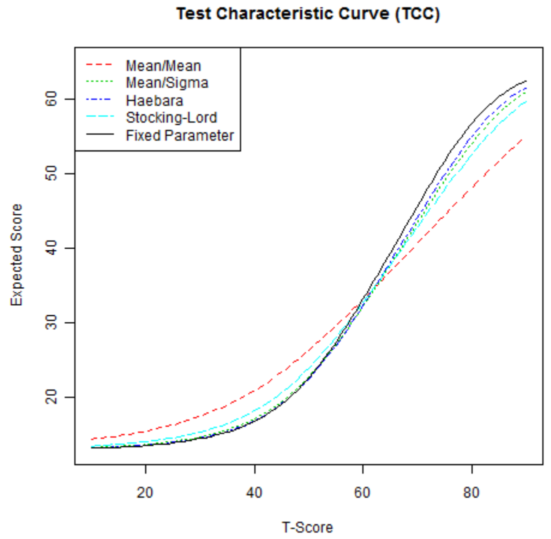


Figure 5.5.5: Test Characteristic Curves (TCC) from Different Linking Methods

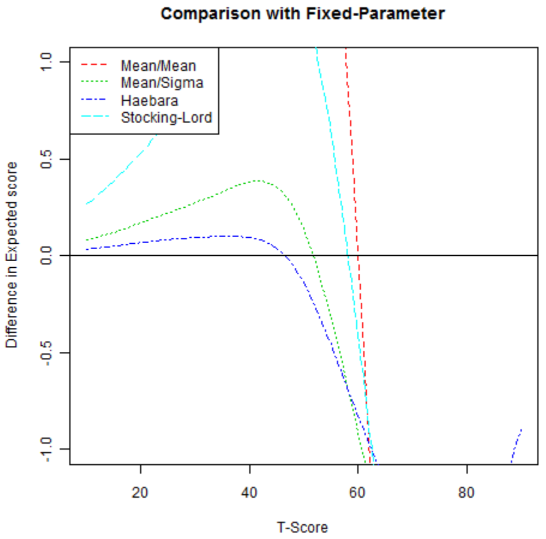


Figure 5.5.6: Difference in Test Characteristic Curves (TCC)

Table 5.5.4 shows the fixed-parameter calibration item parameter estimates for Peds FACIT-F. The marginal reliability estimate for Peds FACIT-F based on the item parameter estimates was 0.899. The marginal reliability estimates for PROMIS Pediatric Fatigue and the combined set were 0.873 and 0.94, respectively. The slope parameter estimates for Peds FACIT-F ranged from 1.07 to 2.64 with a mean of 2.06. The slope parameter estimates for PROMIS Pediatric Fatigue ranged from 0.91 to 1.9 with a mean of 1.4. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.5.7 displays the scale information functions for PROMIS Pediatric Fatigue, Peds FACIT-F, and the combined set of 28. 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 for Peds FACIT-F

a	cb1	cb2	cb3	cb4	NCAT
1.189	-1.539	0.977	2.263	3.442	5
1.067	-0.676	1.623	2.867	4.333	5
1.430	-1.999	0.142	1.731	2.893	5
2.158	-0.682	0.696	1.801	2.596	5
2.059	-0.805	0.635	1.812	3.175	5
2.039	-0.308	0.741	1.764	2.715	5
2.318	-0.080	0.883	1.801	2.723	5
2.558	0.063	0.883	1.885	2.768	5
1.885	0.007	1.132	2.251	3.469	5
2.636	0.291	1.187	2.061	2.523	5
2.422	0.652	1.424	2.288	3.194	5
2.592	0.339	1.049	1.872	2.598	5
2.445	0.543	1.197	1.921	2.759	5

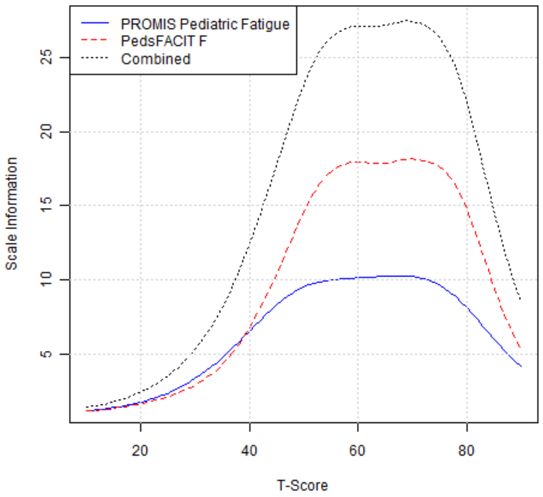


Figure 5.5.7: Comparison of Scale Information Functions

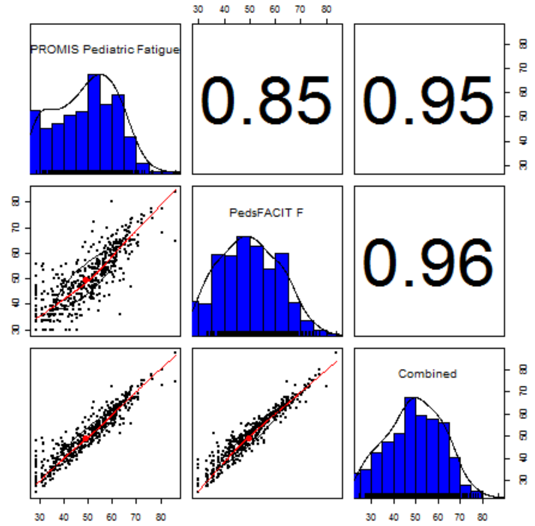


Figure 5.5.8: Comparison of IRT Scaled Scores

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 Peds FACIT-F to a scaled score on PROMIS Pediatric Fatigue can be useful. Based on the Peds FACIT-F 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 Peds FACIT-F to T-score values linked to the PROMIS Pediatric 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.

5.5.6. Equipercentile Linking

We mapped each raw summed score point on Peds FACIT-F to a corresponding scaled score on PROMIS Pediatric Fatigue by identifying scores on PROMIS Pediatric Fatigue that have the same percentile ranks as scores on Peds FACIT-F. Theoretically, the equipercentile linking function is symmetrical for continuous random variables (X and Y). Therefore, the linking function for the values in X to those in Y is the same as that for the values in Y to those in X. However, for discrete variables like raw summed scores the equipercentile linking functions can be slightly different (due to rounding errors and differences in score ranges) and hence may need to be obtained separately. Figure 5.5.9 displays the cumulative distribution functions of the measures. Figure 5.5.10 shows the equipercentile linking functions based on raw summed scores, from Peds FACIT-F to PROMIS Pediatric 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 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).

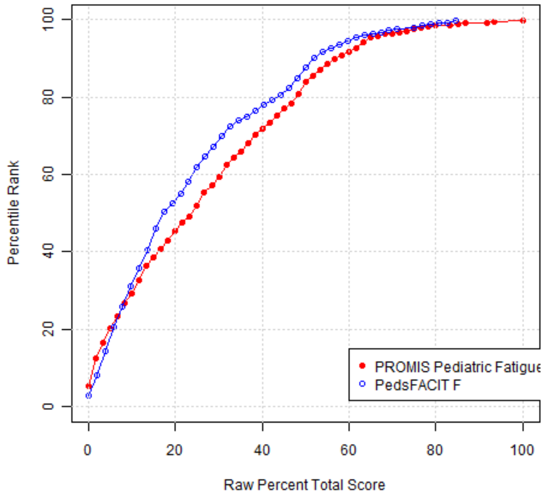


Figure 5.5.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

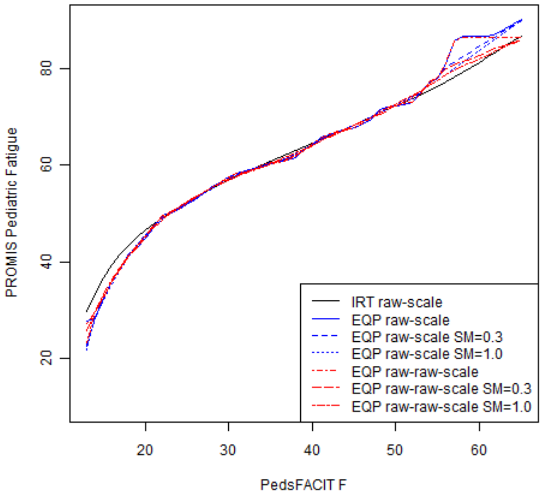


Figure 5.5.10: Equipercentile Linking Functions

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 Pediatric Fatigue T-scores and Peds FACIT-F 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 (0.845), followed by IRT raw-scale (0.839). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (6.909), followed by IRT raw-scale (7.024).

Table 5.5.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.845	-0.735	6.877	6.909
IRT raw-scale	0.839	-0.775	6.988	7.024
EQP raw-scale SM=0.0	0.838	0.597	7.273	7.291
EQP raw-scale SM=0.3	0.836	0.877	7.471	7.515
EQP raw-scale SM=1.0	0.836	0.957	7.513	7.566
EQP raw-raw-scale SM=0.0	0.838	0.503	7.245	7.256
EQP raw-raw-scale SM=0.3	0.837	0.635	7.299	7.320
EQP raw-raw-scale SM=1.0	0.836	0.715	7.367	7.394

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=507) 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 Pediatric 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, 0.734. That is, the means difference between the mean PROMIS Pediatric Fatigue T-score and the mean equated Peds FACIT-F T-score based on a similar sample of 75 cases is expected to be around ±1.47 (i.e., 2 × 0.734).

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	-0.753	1.343	-0.735	0.919	-0.741	0.734
IRT raw-scale	-0.775	1.349	-0.779	0.943	-0.775	0.748
EQP raw-scale SM=0.0	0.630	1.416	0.593	0.975	0.593	0.773
EQP raw-scale SM=0.3	0.895	1.430	0.877	1.006	0.875	0.793
EQP raw-scale SM=1.0	0.940	1.453	0.951	1.018	0.967	0.810
EQP raw-raw-scale SM=0.0	0.486	1.389	0.519	0.969	0.509	0.772
EQP raw-raw-scale SM=0.3	0.661	1.439	0.636	0.976	0.624	0.773
EQP raw-raw-scale SM=1.0	0.693	1.441	0.717	0.990	0.723	0.783

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 are 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 Pediatric Mobility and Neuro-QoL Pediatric Mobility

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of physical function, namely, the PROMIS Pediatric Mobility item bank (eight items) and Neuro-QoL Pediatric Mobility (31 items). Both measures were scaled so that higher scores represent higher levels of physical function. We created raw summed scores for each of the measures separately and then for them 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 503 participants (N = 463 for participants with complete responses).

5.6.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 31 for PROMIS Pediatric Mobility and 128 for Neuro-QoL Pediatric Mobility. 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 them 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 Pediatric Mobility and Neuro-QoL Pediatric Mobility was 0.93. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pediatric Mobility and Neuro-QoL Pediatric Mobility was 0.98. The correlations between the combined score and the measures were 0.95 and 1 for PROMIS Pediatric Mobility and Neuro-QoL Pediatric Mobility, respectively.

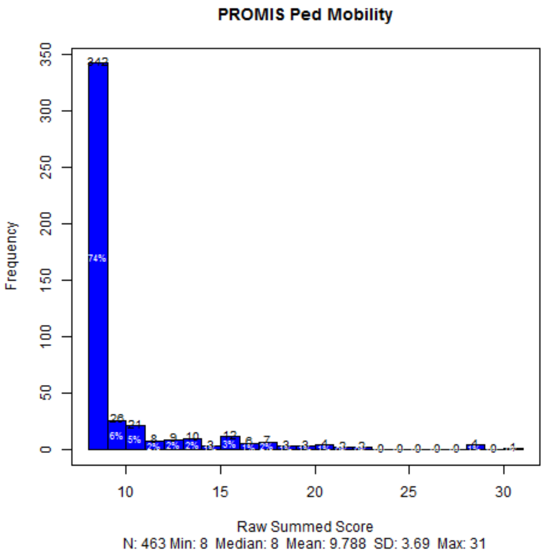


Figure 5.6.1: Raw Summed Score Distribution - PROMIS Pediatric Mobility

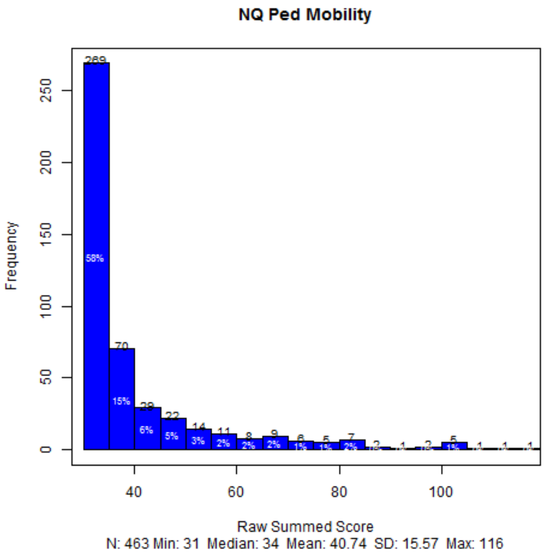


Figure 5.6.2: Raw Summed Score Distribution – Neuro-QoL Pediatric Mobility

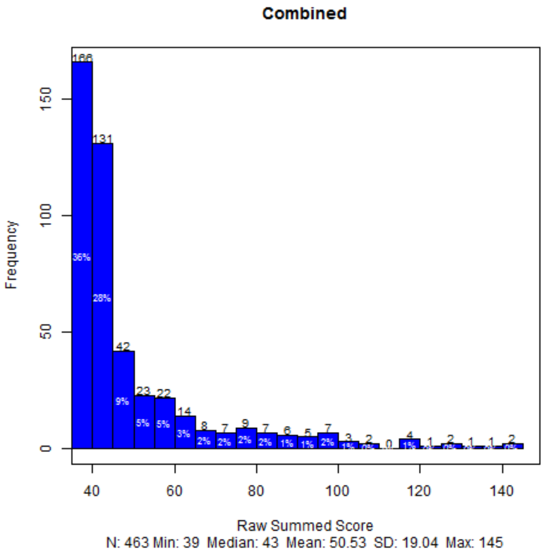


Figure 5.6.3: Raw Summed Score Distribution – Combined

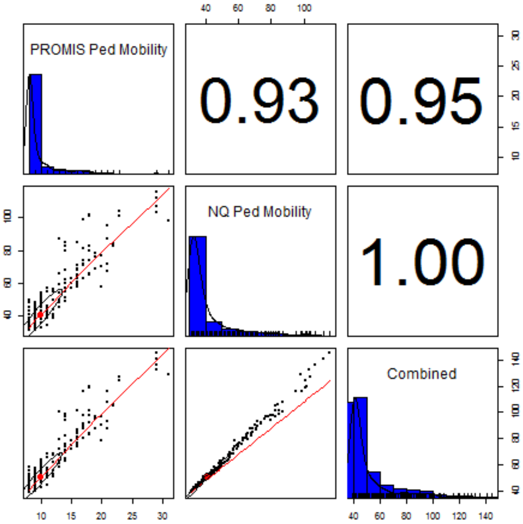


Figure 5.6.4: Scatter Plot Matrix of Raw Summed Scores

5.6.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on them combined. Table 5.6.1 summarizes the results. For PROMIS Pediatric Mobility, Cronbach's alpha internal consistency reliability estimate was 0.922 and adjusted (corrected for overlap) item-total correlations ranged from 0.709 to 0.856. For Neuro-QoL Pediatric Mobility, alpha was 0.972 and adjusted item-total correlations ranged from 0.518 to 0.866. For the 39 items total, alpha was 0.978 and adjusted item-total correlations ranged from 0.519 to 0.863.

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 Pediatric Mobility	8	0.922	0.709	0.758	0.856
Neuro-QoL Pediatric Mobility	31	0.972	0.518	0.760	0.866
Combined	39	0.978	0.519	0.763	0.863

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 them combined. Table 5.6.2 summarizes the model fit statistics. For PROMIS Pediatric Mobility, the fit statistics were as follows: CFI = 0.996, TLI = 0.994, and RMSEA = 0.059. For Neuro-QoL Pediatric Mobility, CFI = 0.984, TLI = 0.983, and RMSEA = 0.067. For the 39 items total, CFI = 0.986, TLI = 0.985, and RMSEA= 0.055. 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 Pediatric Mobility	8	503	0.996	0.994	0.059
Neuro-QoL Pediatric Mobility	31	503	0.984	0.983	0.067
Combined	39	503	0.986	0.985	0.055

5.6.4. Equipercentile Linking

We mapped each raw summed score point on Neuro-QoL Pediatric Mobility to a corresponding scaled score on PROMIS Pediatric Mobility by identifying scores on PROMIS Pediatric Mobility that have the same percentile ranks as scores on Neuro-QoL Pediatric Mobility. Theoretically, the equipercentile linking function is symmetrical for continuous random variables (X and Y). Therefore, the linking function for the values in X to those in Y is the same as that for the values in Y to those in X. However, for discrete variables like raw summed scores, the equipercentile linking functions can be slightly different (due to rounding errors and differences in score ranges) and hence may need to be obtained separately. Figure 5.6.9 displays the cumulative distribution functions of the measures. Figure 5.6.10 shows the equipercentile linking functions based on raw summed scores. 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 16 and Appendix Table 17 show the equipercentile crosswalk tables. The result shown in Appendix Table 16 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 17 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 are 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).

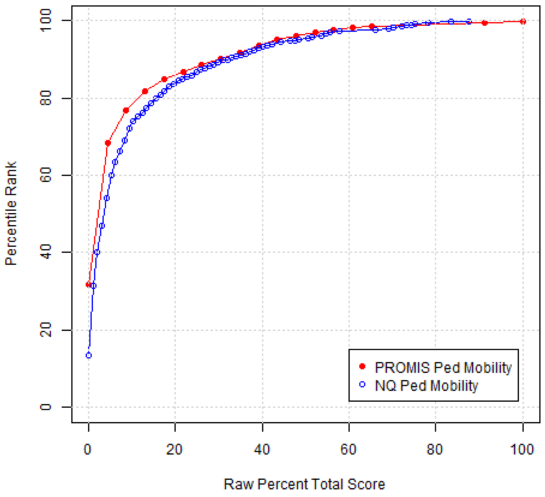


Figure 5.6.5: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

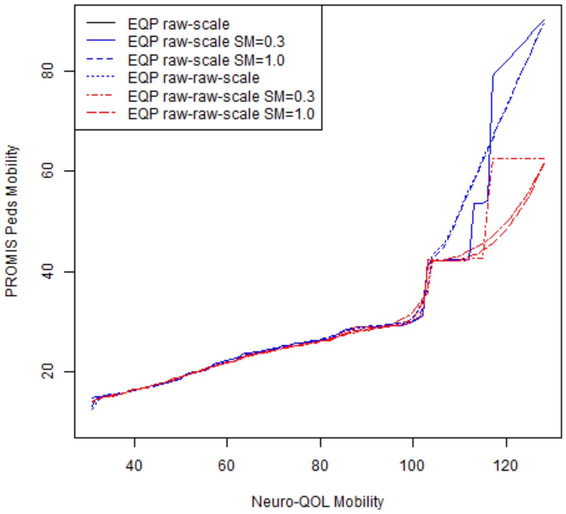


Figure 5.6.10: Equipercentile Linking Functions

5.6.5. 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.

To further facilitate the comparison of the linking methods, Table 5.6.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Pediatric Mobility T-scores and Neuro-QoL Pediatric Mobility scores linked to the T-score metric through different methods. With respect to the correlation between observed and linked T-scores, EQP raw-raw-scale SM=0.3 produced the best result (0.897), followed by EQP raw-raw-scale SM=1.0 (0.895). 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 the smallest RMSD (1.985), followed by EQP raw-raw-scale SM=1.0 (2.007).

Table 5.6.3: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
EQP raw-scale SM=0.0	0.893	-0.516	2.016	2.079
EQP raw-scale SM=0.3	0.872	-0.087	2.569	2.568
EQP raw-scale SM=1.0	0.867	0.076	2.667	2.665
EQP raw-raw-scale SM=0.0	0.892	-0.254	2.063	2.077
EQP raw-raw-scale SM=0.3	0.897	-0.243	1.973	1.985
EQP raw-raw-scale SM=1.0	0.895	-0.191	2.000	2.007

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=463) 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 Pediatric Mobility 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, 0.208. That is, the difference between the mean PROMIS Pediatric Mobility T-score and the mean equated Neuro-QOL Pediatric Mobility T-score based on a similar sample of 75 cases is expected to be around ±0.42 (i.e., 2 × 0.208).

Table 5.6.4: Comparison of Resampling Results

Methods	Mean (N=25)	SD (N=25)	Mean (N=50)	SD (N=50)	Mean (N=75)	SD (N=75)
EQP raw-scale SM=0.0	-0.518	0.395	-0.523	0.270	-0.517	0.212
EQP raw-scale SM=0.3	-0.081	0.494	-0.083	0.344	-0.084	0.272
EQP raw-scale SM=1.0	0.078	0.516	0.075	0.356	0.075	0.282
EQP raw-raw-scale SM=0.0	-0.249	0.399	-0.256	0.279	-0.254	0.218
EQP raw-raw-scale SM=0.3	-0.246	0.373	-0.239	0.262	-0.239	0.211
EQP raw-raw-scale SM=1.0	-0.185	0.396	-0.191	0.261	-0.191	0.208

5.7. PROMIS Pediatric Peer Relationships and Neuro-QoL Pediatric Interaction with Peers

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of social health, namely, the PROMIS Pediatric Peer Relationships item bank (10 items) and Neuro-QoL Pediatric Interaction with Peers (16 items). Both instruments were scaled so that higher scores represent higher levels of social health. We created raw summed scores for each of the measures separately and then for them 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 507 participants (N = 505 for participants with complete responses).

5.7.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 50 for PROMIS Pediatric Peer Relationships and 80 for Neuro-QoL Pediatric Interaction with Peers. 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 them 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 Pediatric Peer Relationships and Neuro-QoL Pediatric Interaction with Peers was 0.85. The disattenuated (corrected for unreliabilities) correlation between PROMIS Pediatric Peer Relationships and Neuro-QoL Pediatric Interaction with Peers was 0.91. The correlations between the combined score and the measures were 0.94 and 0.98 for PROMIS Pediatric Peer Relationships and Neuro-QoL Pediatric Interaction with Peers, respectively.

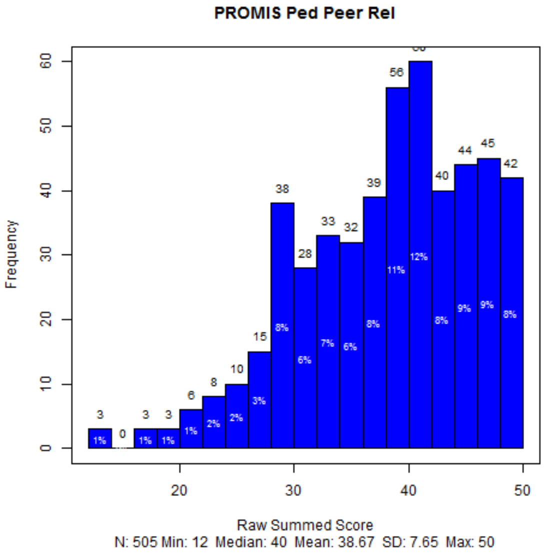


Figure 5.7.1: Raw Summed Score Distribution - PROMIS Pediatric Peer Relationships

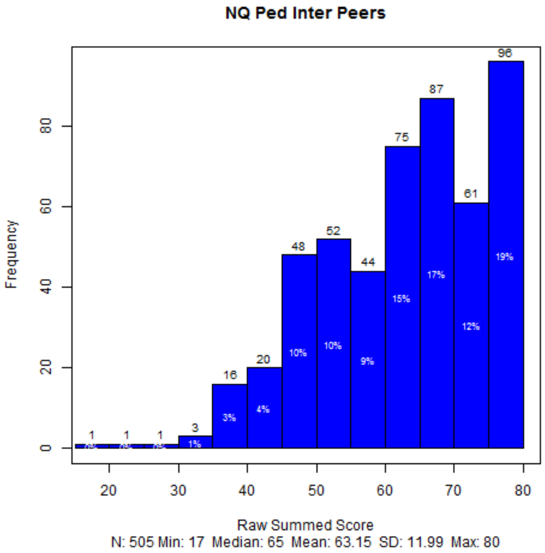


Figure 5.7.2: Raw Summed Score Distribution – Neuro-QoL Pediatric Interaction with Peers

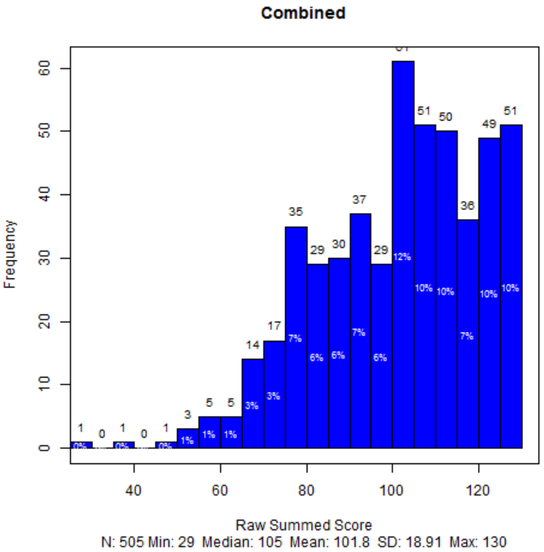


Figure 5.7.3: Raw Summed Score Distribution – Combined

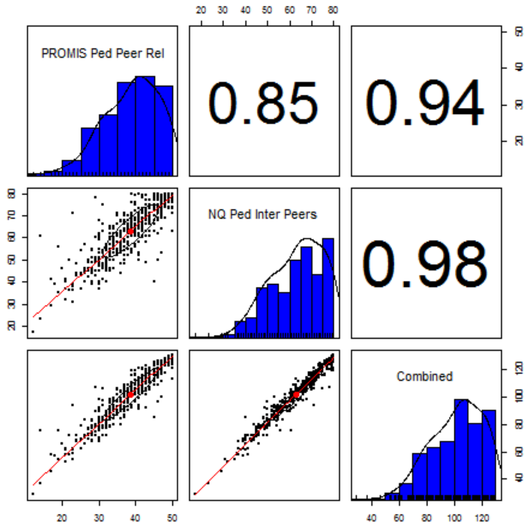


Figure 5.7.4: Scatter Plot Matrix of Raw Summed Scores

5.7.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on them combined. Table 5.7.1 summarizes the results. For PROMIS Pediatric Peer Relationships, Cronbach's alpha internal consistency reliability estimate was 0.923 and adjusted (corrected for overlap) item-total correlations ranged from 0.443 to 0.789. For Neuro-QoL Pediatric Interaction with Peers, alpha was 0.947 and adjusted item-total correlations ranged from 0.484 to 0.807. For the 26 items total, alpha was 0.964 and adjusted item-total correlations ranged from 0.485 to 0.796.

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 Pediatric Peer Relationships	10	0.923	0.443	0.708	0.789
Neuro-QoL Pediatric Interaction with Peers	16	0.947	0.484	0.712	0.807
Combined	26	0.964	0.485	0.705	0.796

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 them combined. Table 5.7.2 summarizes the model fit statistics. For PROMIS Pediatric Peer Relationships, the fit statistics were as follows: CFI = 0.971, TLI =

0.963, and RMSEA = 0.132. For Neuro-QoL Pediatric Interaction with Peers, CFI = 0.972, TLI = 0.968, and RMSEA = 0.101. For the 26 items total, CFI = 0.946, TLI = 0.941, and RMSEA = 0.101. 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 Pediatric Peer Relationships	10	507	0.971	0.963	0.132
Neuro-QoL Pediatric Interaction with Peers	16	507	0.972	0.968	0.101
Combined	26	507	0.946	0.941	0.101

5.7.4. 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 10 PROMIS Pediatric Peer Relationships items served as anchor items to transform the item parameter estimates for the Neuro-QoL Pediatric Interaction with Peers items onto the PROMIS Pediatric Peer Relationships 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 Pediatric Peer Relationships items were constrained to their final bank values, while the Neuro-QoL Pediatric Interaction with Peers items were calibrated under the constraints imposed by the anchor items.

Table 5.7.3: IRT Linking Constants

	A	B
Mean/Mean	1.412	-0.362
Mean/Sigma	0.966	-0.790
Haebara	0.909	-0.740
Stocking-Lord	1.061	-0.659

The item parameter estimates for the Neuro-QoL Pediatric Interaction with Peers items were linked to the PROMIS Pediatric Peer Relationship metric using the transformation constants shown in Table 5.7.3. The Neuro-QoL Pediatric Interaction with Peers item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Pediatric Peer Relationships metric. Based on the transformed and fixed-parameter estimates, we derived test characteristic curves (TCC) for Neuro-QoL Pediatric Interaction with Peers 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.

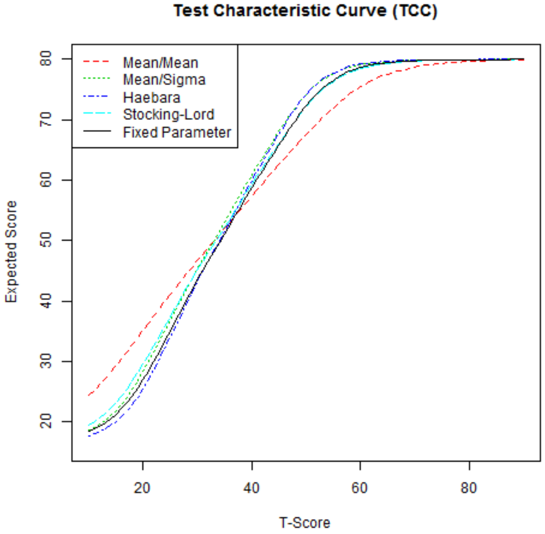


Figure 5.7.5: Test Characteristic Curves (TCC) from Different Linking Methods

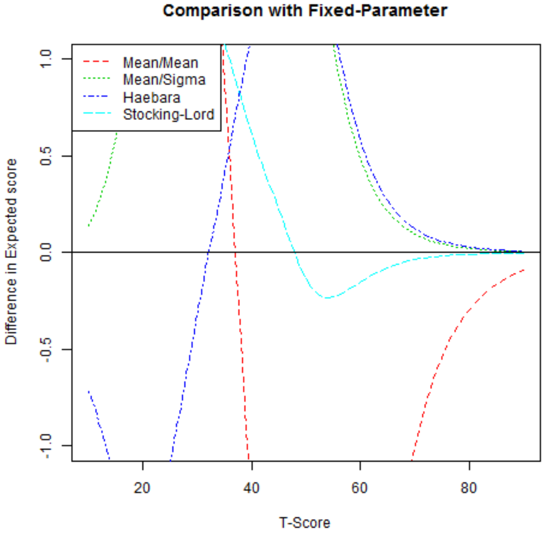


Figure 5.7.6: Difference in Test Characteristic Curves (TCC)

Table 5.7.4 shows the fixed-parameter calibration item parameter estimates for Neuro-QoL Pediatric Interaction with Peers. The marginal reliability estimate for Neuro-QoL Pediatric Interaction with Peers based on the item parameter estimates was 0.897. The marginal reliability estimates for PROMIS Pediatric Peer Relationships and the combined set were 0.822 and 0.927, respectively. The slope parameter estimates for Neuro-QoL Pediatric Interaction with Peers ranged from 1.26 to 3.6, with a mean of 2.56. The slope parameter estimates for PROMIS Pediatric Peer Relationships ranged from 0.65 to 2.06, with a mean of 1.66. We also derived scale information functions based on the fixed- parameter calibration result. Figure 5.7.7 displays the scale information functions for PROMIS Pediatric Peer Relationships, Neuro-QoL Pediatric Interaction with Peers, and the combined set of 26 items. 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 Neuro-QoL Pediatric Interaction with Peers

a	cb1	cb2	cb3	cb4	NCAT
1.444	-2.820	-2.012	-1.109	-0.131	5
1.258	-4.343	-3.236	-1.607	-0.114	5
1.993	-2.920	-2.095	-1.031	0.263	5
2.419	-2.959	-2.096	-0.943	0.218	5
2.862	-2.746	-2.192	-1.202	-0.215	5
3.167	-2.866	-2.262	-1.326	-0.361	5
3.161	-3.028	-2.326	-1.339	-0.361	5
2.831	-2.859	-2.221	-1.235	-0.143	5
2.747	-2.884	-2.134	-1.136	-0.059	5
3.600	-2.950	-2.289	-1.382	-0.365	5

3.222	-2.895	-2.019	-1.137	-0.170	5
3.104	-3.066	-2.094	-1.100	-0.141	5
2.210	-3.210	-2.327	-1.037	0.007	5
1.834	-3.392	-2.286	-1.132	-0.087	5
2.395	-3.398	-2.522	-1.353	-0.272	5
2.773	-2.917	-2.282	-1.124	-0.034	5

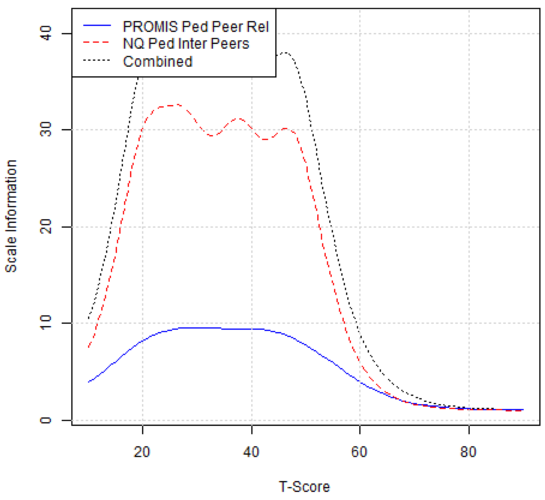


Figure 5.7.7: Comparison of Scale Information Functions

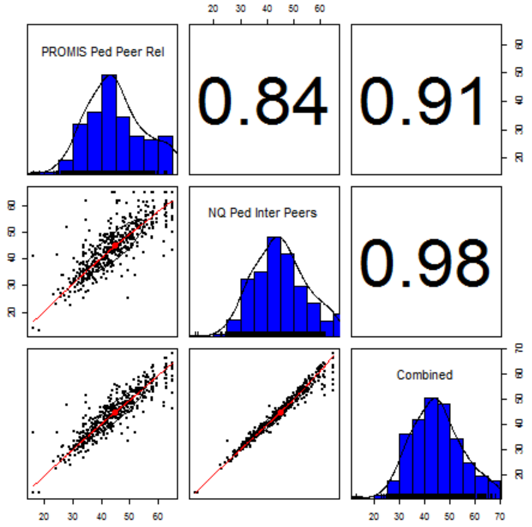


Figure 5.7.8: Comparison of IRT Scaled Scores

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 from Neuro-QoL Pediatric Interaction with Peers to a scaled score on PROMIS Pediatric Peer Relationships can be useful. Based on the Neuro-QoL Pediatric Interaction with Peer item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Appendix Table 18 can be used to map simple raw summed scores Neuro-QoL Pediatric Interaction with Peers to T-score values linked to the PROMIS Pediatric Peer Relationships 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 so 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 Neuro-QoL Pediatric Interaction with Peers to a corresponding scaled score on PROMIS Pediatric Peer Relationships by identifying scores on

PROMIS Pediatric Peer Relationships that have the same percentile ranks as scores on Neuro-QoL Pediatric Interaction with Peers. 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, Neuro-QoL Pediatric Interaction with Peers to PROMIS Pediatric Peer Relationships. 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 19 and Appendix Table 20 show the equipercentile crosswalk tables. The result shown in Appendix Table 19 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 20 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 are 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).

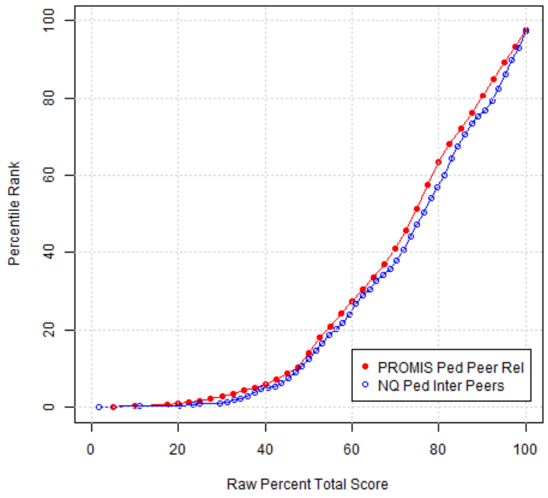


Figure 5.7.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

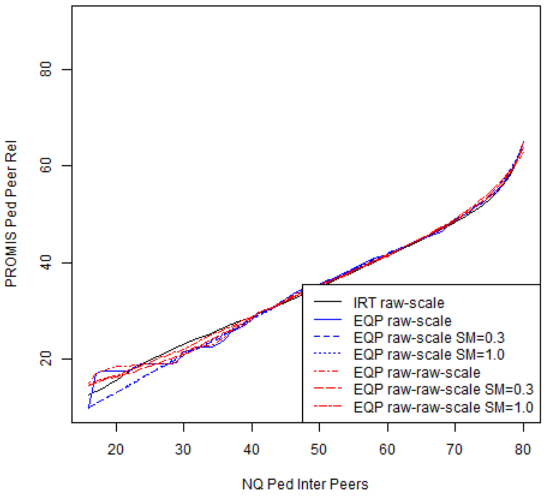


Figure 5.7.10: Equipercentile Linking Functions

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. Another potential issue with IRT-based linking methods is the violation of model assumptions as a result of combining items from two measures (e.g., unidimensionality and local independence). As displayed in Figure 5.7.10, the relationships derived from various linking methods are consistent, which suggests that a robust linking relationship can be determined based on the given sample.

To further facilitate the comparison of the linking methods, Table 5.7.5 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Pediatric Peer Relationships T-scores and Neuro-QoL Pediatric Interaction with Peers 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 (0.839), followed by EQP raw-raw-scale SM=1.0 (0.838). 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 the smallest RMSD (5.634), followed by EQP raw-raw-scale SM=0.3 (5.647)

Table 5.7.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.839	-0.061	5.676	5.670
IRT raw-scale	0.835	0.022	5.674	5.668
EQP raw-scale SM=0.0	0.835	-0.317	5.683	5.686
EQP raw-scale SM=0.3	0.835	-0.320	5.688	5.692
EQP raw-scale SM=1.0	0.836	-0.328	5.687	5.690
EQP raw-raw-scale SM=0.0	0.836	-0.250	5.679	5.679
EQP raw-raw-scale SM=0.3	0.837	-0.269	5.646	5.647
EQP raw-raw-scale SM=1.0	0.838	-0.283	5.632	5.634

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=505) 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 Pediatric Peer Relationships 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 and EQP raw-raw-scale SM=1.0 produced the smallest standard error, 0.6. That is, the difference between the mean PROMIS Pediatric Peer Relationships T-score and the mean equated Neuro-QoL Pediatric Interaction with Peers T-score based on a similar sample of 75 cases is expected to be around ±1.2 (i.e., 2 × 0.6).

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	-0.065	1.100	-0.064	0.763	-0.058	0.608
IRT raw-scale	0.014	1.097	0.028	0.759	0.021	0.600
EQP raw-scale SM=0.0	-0.317	1.107	-0.314	0.745	-0.315	0.606
EQP raw-scale SM=0.3	-0.324	1.110	-0.324	0.764	-0.318	0.605
EQP raw-scale SM=1.0	-0.340	1.099	-0.325	0.766	-0.328	0.608
EQP raw-raw-scale SM=0.0	-0.252	1.100	-0.251	0.771	-0.261	0.603
EQP raw-raw-scale SM=0.3	-0.274	1.092	-0.264	0.760	-0.272	0.607
EQP raw-raw-scale SM=1.0	-0.271	1.082	-0.280	0.756	-0.281	0.600

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 are 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 18) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

5.8. Neuro-QoL Pediatric Cognitive Function and Pediatric Perceived Cognitive Function (Peds PCF)

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of cognition, namely the Neuro-QoL Pediatric Cognitive Function (NQ Peds Cog) item bank (14 items) and Peds PCF (30 items). Both instruments were scaled so that higher scores represent higher levels of cognition. We created raw summed scores for each of the measures separately and then for them 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 507 participants (N = 505 for participants with complete responses).

5.8.1. Raw Summed Score Distribution

The maximum possible raw summed scores were 70 for NQ Peds Cog and 150 for PedsPCF. Figure 5.8.1 and Figure 5.8.2 graphically display the raw summed score of the two measures. Figure 5.8.3 shows the distribution for them 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 NQ Peds Cog and PedsPCF was 0.93. The disattenuated (corrected for unreliabilies) correlation between NQ Peds Cog and Peds PCF was 0.96. The correlations between the combined score and the measures were 0.97 and 0.99 for NQ Peds Cog and PedsPCF, respectively.

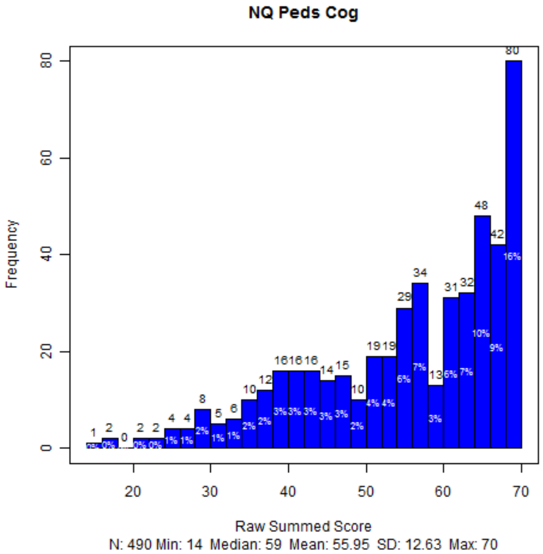


Figure 5.8.1: Raw Summed Score Distribution - Neuro-QoL Peds Cognitive Function

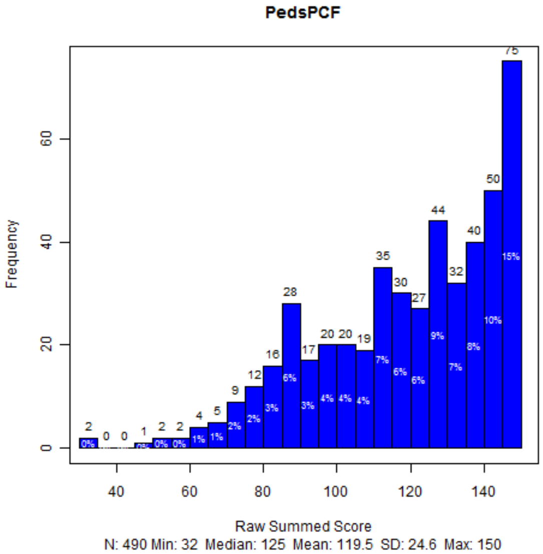


Figure 5.8.2: Raw Summed Score Distribution – Peds PCF

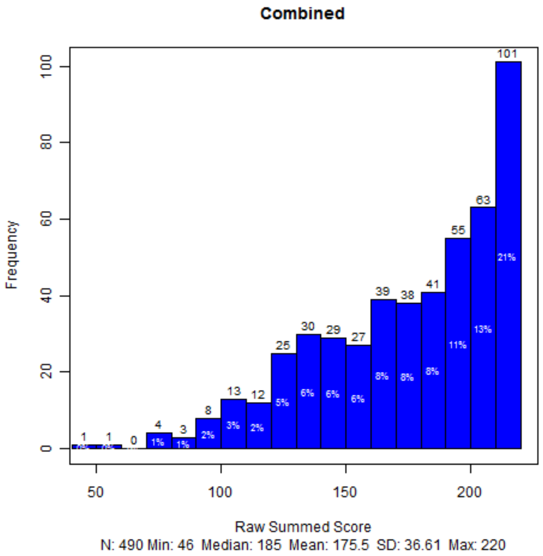


Figure 5.8.3: Raw Summed Score Distribution – Combined

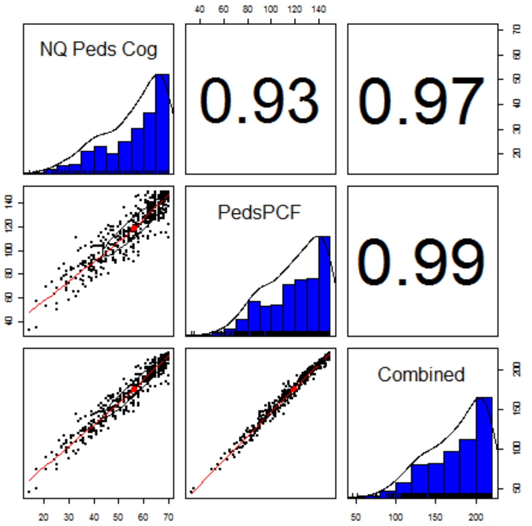


Figure 5.8.4: Scatter Plot Matrix of Raw Summed Scores

5.8.2. Classical Item Analysis

We conducted classical item analyses on the two measures separately and on them combined. Table 5.8.1 summarizes the results. For NQ Peds Cog, Cronbach's alpha internal consistency reliability estimate was 0.959 and adjusted (corrected for overlap) item-total correlations ranged from 0.712 to 0.833. For Peds PCF, alpha was 0.975 and adjusted item-total correlations ranged from 0.535 to 0.826. For the 44 items, alpha was 0.983 and adjusted item-total correlations ranged from 0.52 to 0.837.

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
NQ Peds Cog	14	0.959	0.712	0.774	0.833
PedsPCF	30	0.975	0.535	0.740	0.826
Combined	44	0.983	0.520	0.749	0.837

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 NQ Peds Cog, the fit statistics were as follows: CFI = 0.977, TLI= 0.972, and RMSEA = 0.116. For Peds PCF, CFI = 0.963, TLI = 0.960, and RMSEA = 0.081. For the 44 items, CFI = 0.951, TLI = 0.949, and RMSEA = 0.81.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
NQ Peds Cog	14	507	0.977	0.972	0.116
PedsPCF	30	507	0.963	0.960	0.081
Combined	44	507	0.951	0.949	0.081

5.8.4. 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 14 NQ Peds Cog items served as anchor items to transform the item parameter estimates for the Peds PCF items onto the NQ Peds Cog 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 NQ Peds Cog items were constrained to their final bank values, while the Peds PCF items were calibrated, under the constraints imposed by the anchor items.

Table 5.8.3: IRT Linking Constants

	A	B
Mean/Mean	1.010	-0.322
Mean/Sigma	1.060	-0.290
Haebara	1.054	-0.294
Stocking-Lord	1.048	-0.298

The item parameter estimates for the PedsPCF items were linked to the NQ Peds Cog metric using the transformation constants shown in Table 5.8.3. The Peds PCF item parameter estimates from the fixed-parameter calibration are considered already on the NQ Peds Cog metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for Peds PCF 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.

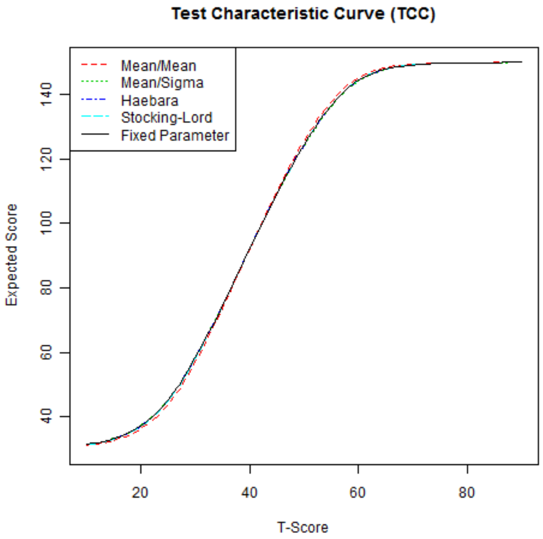


Figure 5.8.5: Test Characteristic Curves (TCC) from Different Linking Methods

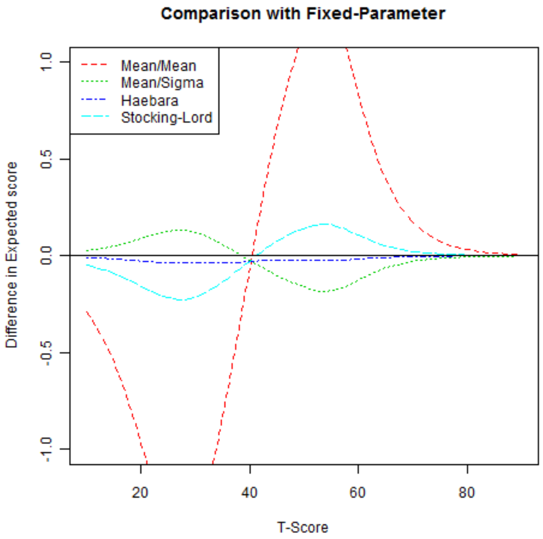


Figure 5.8.6: Difference in Test Characteristic Curves (TCC)

Table 5.8.4 shows the fixed-parameter calibration item parameter estimates for Peds PCF. The marginal reliability estimate for PedsPCF based on the item parameter estimates was 0.96. The marginal reliability estimates for NQ Peds Cog and the combined set were 0.926 and 0.971, respectively. The slope parameter estimates for PedsPCF ranged from 1.37 to 3.61 with a mean of 2.51. The slope parameter estimates for NQ Peds Cog ranged from 2.18 to 3.74 with a mean of 2.89. We also derived scale information functions based on the fixed-parameter calibration result. Figure 5.8.7 displays the scale information functions for NQ Peds Cog, PedsPCF, and the combined set of 44. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 5.8.8 is a scatter plot matrix showing the relationships between the measures.

Table 5.8.4: Fixed-Parameter Calibration Item Parameter Estimates for Peds PCF

a	cb1	cb2	cb3	cb4	NCAT
1.366	-3.753	-2.569	-1.406	-0.270	5
2.056	-2.399	-1.528	-0.466	0.539	5
1.827	-2.945	-1.849	-0.831	0.304	5
2.486	-2.005	-1.352	-0.507	0.418	5
2.648	-2.289	-1.547	-0.721	0.259	5
2.182	-2.622	-1.728	-1.008	-0.208	5
2.781	-2.220	-1.355	-0.629	0.255	5
1.754	-1.661	-0.798	0.115	1.167	5
2.345	-2.468	-1.722	-0.916	0.099	5
2.712	-2.287	-1.402	-0.566	0.330	5
2.427	-2.235	-1.340	-0.532	0.467	5
2.328	-2.293	-1.493	-0.511	0.608	5
2.196	-2.607	-1.612	-0.733	0.329	5
2.142	-2.916	-1.971	-1.104	-0.312	5
2.159	-2.147	-1.466	-0.837	0.040	5
1.681	-3.109	-2.253	-1.172	-0.207	5
2.148	-2.176	-1.269	-0.391	0.719	5
2.805	-1.902	-1.184	-0.453	0.474	5
2.488	-2.305	-1.355	-0.531	0.491	5
2.496	-2.515	-1.797	-0.806	0.242	5
2.734	-2.219	-1.317	-0.545	0.451	5
2.548	-2.567	-1.592	-0.906	-0.202	5
2.796	-2.349	-1.452	-0.625	0.247	5
3.050	-1.828	-1.244	-0.555	0.206	5
3.218	-2.011	-1.480	-0.716	0.100	5
2.409	-2.386	-1.597	-0.939	-0.190	5
3.238	-2.001	-1.394	-0.649	0.257	5

3.314	-1.861	-1.124	-0.453	0.466	5
3.609	-1.822	-1.153	-0.431	0.420	5
3.480	-1.939	-1.145	-0.536	0.311	5

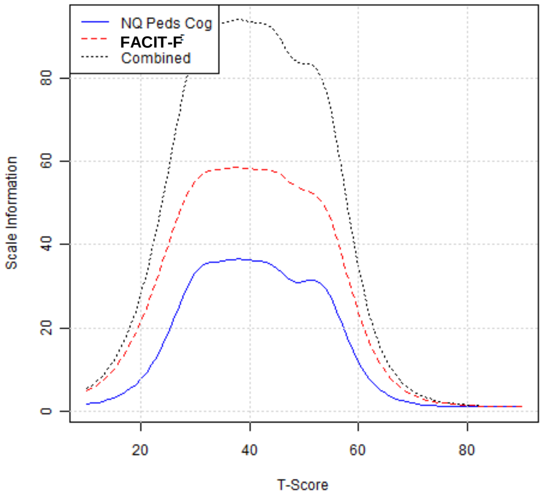


Figure 5.8.7: Comparison of Scale Information Functions

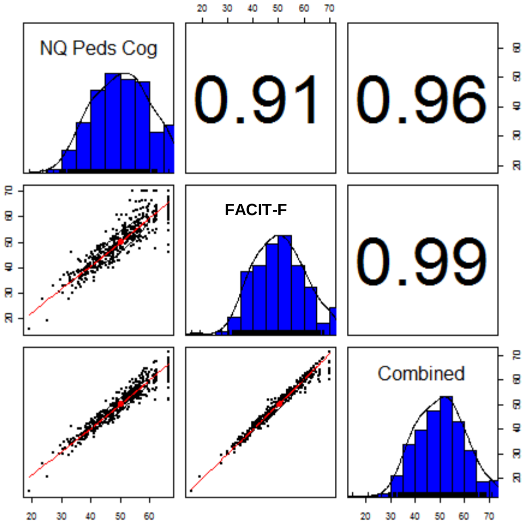


Figure 5.8.8: Comparison of IRT Scaled Scores

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 Peds PCF to a scaled score on NQ Peds Cog 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 21 can be used to map simple raw summed scores from Peds PCF to T-score values linked to the NQ Peds Cog metric. Each raw summed score point and corresponding NQ Peds Cog 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 PedsPCF to a corresponding scaled score on NQ Peds Cog by identifying scores on NQ Peds Cog 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 5.8.9 displays the cumulative distribution functions of the measures. Figure 5.8.10 shows the equipercentile linking functions based on raw summed scores, from Peds PCF to NQ Peds Cog. 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 22 and Appendix Table 23 show the equipercentile crosswalk tables. The result shown in Appendix Table 22 is based on the direct (raw summed score to scaled score) approach, whereas Appendix Table 23 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).

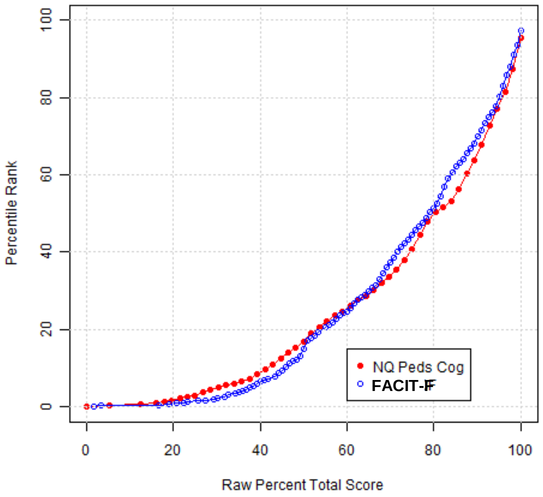


Figure 5.8.9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

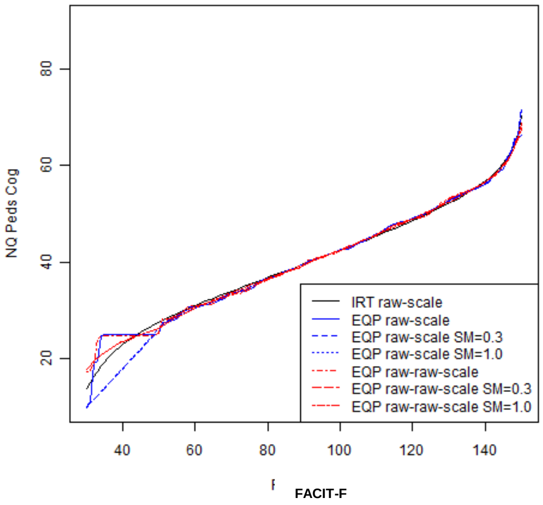


Figure 5.8.10: Equipercentile Linking Functions

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 NQ Peds Cog 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 5.8.10), 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 (0.911), followed by EQP raw-raw-scale SM=0.3 (0.91). 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 (4.067), followed by EQP raw-raw- scale SM=0.3 (4.071).

Table 5.8.5: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.911	-0.079	4.143	4.140
IRT raw-scale	0.903	-0.176	4.320	4.320
EQP raw-scale SM=0.0	0.909	0.009	4.071	4.067
EQP raw-scale SM=0.3	0.901	-0.245	4.424	4.426
EQP raw-scale SM=1.0	0.902	-0.266	4.379	4.382
EQP raw-raw-scale SM=0.0	0.909	-0.052	4.091	4.087
EQP raw-raw-scale SM=0.3	0.910	0.015	4.075	4.071
EQP raw-raw-scale SM=1.0	0.907	-0.029	4.172	4.168

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=490) 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 NQ Peds Cog 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, 0.432. That is, the difference between the mean NQ Peds Cog T-score and the mean equated PedsPCF T-score based on a similar sample of 75 cases is expected to be around ±0.86 (i.e., 2 × 0.432).

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	-0.083	0.804	-0.084	0.548	-0.069	0.438
IRT raw-scale	-0.172	0.844	-0.171	0.580	-0.179	0.461
EQP raw-scale SM=0.0	0.010	0.797	0.002	0.544	0.009	0.435
EQP raw-scale SM=0.3	-0.233	0.863	-0.248	0.596	-0.237	0.468
EQP raw-scale SM=1.0	-0.277	0.854	-0.266	0.586	-0.268	0.465
EQP raw-raw-scale SM=0.0	-0.058	0.794	-0.059	0.546	-0.056	0.432
EQP raw-raw-scale SM=0.3	0.007	0.788	0.019	0.545	0.012	0.438
EQP raw-raw-scale SM=1.0	-0.030	0.813	-0.034	0.561	-0.028	0.447

Examining a number of linking studies in the current project revealed that the two linking methods (IRT and equipercentile) in general produced highly comparable results. Some noticeable discrepancies were observed (albeit rarely) in some extreme score levels where data were sparse. Model-based approaches can provide more robust results than those relying solely on data when data are 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 21) might be preferred when the results are comparable and no apparent violations of assumptions are evident.

6.0 References

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

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

Beck, A. T., Steer, R. A., Ball, R., Ranieri, W. (December 1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of personality assessment 67* NA

Beck, A. T., Steer, R. A. and Brown, G. K. (1996) *Manual for the Beck Depres- sion Inventory-II*. San Antonio, TX.: Psychological Corporation.

Beck, A. T., Ward. C., Mendelson, M. (1961). Beck Depression Inventory (BDI). *Arch Gen Psychiatry 4* NA

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., Erbaugh, J. (June 1961). An inventory for measuring depression. *Arch. Gen. Psychiatry 4* NA

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

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

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

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

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

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

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

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

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

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

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

Faulstich, M.E., Carey, M.P., Ruggiero, L., Enyart, P., & Gresham, F. (1986). Assessment of Depression in Childhood and Adolescence: An Evaluation of the Center for Epidemiological Studies Depression Scale for Children (CES-DC). *American Journal of Psychiatry*, 143(8), 1024-1027.

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

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

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

Lai, J.S., Cella, D., Kupst, M.J., Holm, S., Kelly, M.E., Bode, R.K., & Goldman, S. (2007). Measuring Fatigue for Children with Cancer: Development and Validation of the Pediatric Functional Assessment of Chronic Illness Therapy-Fatigue (PedsFACIT-F). *Journal of Pediatric Hematology/Oncology*, 29(7), 471-479.

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. (In Press). Cognition Reported by Parent of Children with Cancer Compared to It Reported by Parents of Us Pediatric General Population. *Quality of Life Research*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Weissman, M.M., Orvaschel, H., & Padian, N. (1980). Children's Symptom and Social Functioning Self-Report Scales Comparison of Mothers' and Children's Reports. *Journal of Nervous and Mental Disease, 168*(12), 736-740.

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

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 Neuro-QoL Pediatric Anxiety Full Item Bank and PROMIS Pediatric Anxiety (Neuro-QoL Wave 1 Study) - RECOMMENDED

Neuro-QoL Peds Anxiety T-Score	Neuro-QoL Peds Anxiety Raw Score	PROMIS Anxiety T-Score	T-Score SE
35.2	19	31.8	5.4
39.9	20	35.9	4.4
42.3	21	38.2	4.0
44.0	22	40.2	3.6
45.3	23	41.7	3.3
46.5	24	43.1	3.0
47.4	25	44.3	2.8
48.2	26	45.3	2.6
48.9	27	46.3	2.5
49.6	28	47.1	2.3
50.2	29	47.9	2.2
50.8	30	48.7	2.2
51.3	31	49.4	2.1
51.8	32	50.0	2.0
52.3	33	50.6	2.0
52.8	34	51.2	2.0
53.3	35	51.8	1.9
53.7	36	52.4	1.9
54.1	37	52.9	1.9
54.6	38	53.4	1.9
55.0	39	54.0	1.9
55.4	40	54.5	1.8
55.8	41	55.0	1.8
56.2	42	55.5	1.8
56.6	43	55.9	1.8
57.0	44	56.4	1.8
57.4	45	56.9	1.8
57.8	46	57.4	1.8
58.2	47	57.9	1.8
58.6	48	58.3	1.8
59.0	49	58.8	1.8

59.4	50	59.3	1.8
59.8	51	59.7	1.8
60.2	52	60.2	1.8
60.5	53	60.7	1.8
60.9	54	61.1	1.8
61.3	55	61.6	1.9
61.7	56	62.0	1.9
62.1	57	62.5	1.8
62.5	58	62.9	1.8
62.9	59	63.4	1.8
63.2	60	63.8	1.8
63.6	61	64.3	1.8
64.0	62	64.8	1.8
64.4	63	65.2	1.8
64.8	64	65.7	1.8
65.2	65	66.1	1.8
65.5	66	66.6	1.8
65.9	67	67.0	1.8
66.3	68	67.5	1.8
66.7	69	67.9	1.8
67.0	70	68.4	1.8
67.4	71	68.9	1.8
67.8	72	69.3	1.8
68.2	73	69.8	1.8
68.5	74	70.3	1.8
68.9	75	70.7	1.8
69.3	76	71.2	1.8
69.7	77	71.7	1.8
70.1	78	72.2	1.8
70.5	79	72.7	1.8
70.9	80	73.2	1.9
71.3	81	73.7	1.9
71.7	82	74.3	1.9
72.2	83	74.8	2.0
72.6	84	75.4	2.0
73.1	85	76.0	2.0
73.6	86	76.7	2.1
74.2	87	77.4	2.2
74.8	88	78.1	2.3
75.4	89	78.9	2.4
76.2	90	79.8	2.5

77.0	91	80.8	2.6
77.9	92	81.8	2.8
79.1	93	82.9	2.8
80.5	94	84.1	2.8
82.6	95	85.5	2.7

Appendix Table 2: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Pediatric Anxiety Full Item Bank to PROMIS Pediatric Anxiety
Note: Table 1 is recommended

Neuro-QoL Ped Anxiety Raw 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)
19	37	27	26	0.04
20	37	36	34	0.04
21	37	37	37	0.04
22	37	38	38	0.05
23	43	42	42	0.16
24	43	43	43	0.14
25	43	44	44	0.15
26	46	45	45	0.16
27	46	46	46	0.15
28	46	46	46	0.12
29	46	47	47	0.13
30	50	49	49	0.14
31	50	50	50	0.14
32	50	50	50	0.15
33	50	50	51	0.15
34	50	51	51	0.16
35	53	52	52	0.18
36	53	53	53	0.17
37	53	53	53	0.15
38	53	54	54	0.14
39	56	55	55	0.15
40	56	56	55	0.14
41	56	56	56	0.15
42	56	56	56	0.13
43	56	56	56	0.13
44	56	56	56	0.13
45	56	56	56	0.14
46	56	57	57	0.13
47	56	57	57	0.13
48	56	57	57	0.13
49	56	57	57	0.13
50	59	58	58	0.11
51	59	58	58	0.11
52	59	58	58	0.10
53	59	58	59	0.10

54	59	59	59	0.10
55	59	59	59	0.09
56	59	59	59	0.09
57	59	59	59	0.09
58	59	59	59	0.09
59	59	59	59	0.08
60	59	59	59	0.08
61	59	59	60	0.08
62	59	60	60	0.09
63	59	60	60	0.09
64	62	61	61	0.53
65	62	62	62	0.46
66	62	63	63	0.46
67	65	64	63	0.41
68	65	65	64	0.37
69	65	65	65	0.34
70	65	65	65	0.30
71	65	66	66	0.28
72	67	67	67	0.27
73	68	67	67	0.31
74	68	68	68	0.21
75	68	68	68	0.22
76	68	69	69	0.17
77	70	70	70	0.47
78	70	70	70	0.61
79	70	71	71	0.61
80	74	73	73	1.00
81	74	75	74	0.94
82	76	76	75	0.49
83	76	76	76	0.49
84	77	76	76	0.49
85	77	77	77	0.40
86	77	77	77	0.40
87	77	77	78	0.40
88	77	78	78	0.40
89	77	78	79	0.31
90	77	80	81	0.28
91	77	82	83	0.18
92	87	84	84	0.18
93	88	86	86	0.18
94	89	88	88	0.18
95	90	90	90	0.18

Appendix Table 3: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – Neuro-QoL Pediatric Anxiety Full Item Bank to PROMIS Pediatric Anxiety
Note: Table 1 is recommended

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

61	60	60	61
62	60	61	61
63	60	61	62
64	61	62	62
65	62	62	62
66	63	62	63
67	63	63	63
68	64	64	64
69	65	64	64
70	65	65	65
71	66	65	66
72	66	66	66
73	67	67	67
74	68	68	68
75	68	68	68
76	69	69	69
77	70	70	70
78	71	71	70
79	72	72	71
80	73	72	72
81	74	73	72
82	75	74	73
83	75	74	74
84	76	75	74
85	76	75	75
86	76	76	76
87	76	77	77
88	76	78	78
89	77	78	78
90	78	78	79
91	79	78	79
92	79	79	79
93	79	79	79
94	79	79	79
95	79	79	79

Appendix Table 4: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for CES-D Children and PROMIS Pediatric Depressive Symptoms (PROsetta Stone Wave 2 Study) - RECOMMENDED

CES-D Children Actual Score*	CES-D Children Raw Score*	PROMIS Peds Depression T-Score	SE	CES-D Children Actual Score*	CES-D Children Raw Score*	PROMIS Peds Depression T-Score	SE
0	20	31.8	6.0	30	50	64.2	2.4
1	21	34.3	5.7	31	51	64.9	2.4
2	22	37.1	5.2	32	52	65.5	2.4
3	23	39.3	4.8	33	53	66.1	2.4
4	24	41.3	4.4	34	54	66.7	2.4
5	25	43.0	4.1	35	55	67.4	2.4
6	26	44.6	3.8	36	56	68.0	2.4
7	27	46.0	3.5	37	57	68.6	2.4
8	28	47.3	3.3	38	58	69.3	2.4
9	29	48.4	3.1	39	59	69.9	2.4
10	30	49.5	3.0	40	60	70.6	2.4
11	31	50.5	2.9	41	61	71.2	2.4
12	32	51.4	2.8	42	62	71.9	2.5
13	33	52.3	2.7	43	63	72.6	2.5
14	34	53.2	2.7	44	64	73.3	2.5
15	35	54.0	2.6	45	65	74.0	2.6
16	36	54.8	2.6	46	66	74.8	2.6
17	37	55.5	2.6	47	67	75.6	2.7
18	38	56.3	2.5	48	68	76.4	2.8
19	39	57.0	2.5	49	69	77.2	2.9
20	40	57.7	2.5	50	70	78.1	3.0
21	41	58.4	2.5	51	71	79.1	3.0
22	42	59.1	2.5	52	72	80.1	3.1
23	43	59.7	2.4	53	73	81.1	3.2
24	44	60.4	2.4	54	74	82.1	3.2
25	45	61.1	2.4	55	75	83.2	3.2
26	46	61.7	2.4	56	76	84.2	3.1
27	47	62.3	2.4	57	77	85.1	2.9
28	48	63.0	2.4	58	78	85.9	2.7
29	49	63.6	2.4	59	79	86.6	2.4

*The scores in the first column conform to the scoring rules of the CES-D Children instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 5: Direct (Raw to Scale) Equipercentile Crosswalk Table – From CES-D Children to PROMIS Pediatric Depressive Symptoms – Table 4 is recommended

CES-D Children Actual Score*	CES-D Children Raw 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	20	32	25	25	0.06
1	21	32	32	32	0.06
2	22	35	34	34	0.21
3	23	37	37	37	0.21
4	24	39	39	39	0.24
5	25	41	41	41	0.22
6	26	42	42	42	0.30
7	27	44	44	44	0.34
8	28	45	46	46	0.36
9	29	48	48	48	0.43
10	30	49	49	49	0.53
11	31	51	51	50	0.62
12	32	52	52	51	0.22
13	33	52	52	52	0.20
14	34	53	53	53	0.50
15	35	54	54	54	0.42
16	36	55	55	55	0.54
17	37	56	55	55	0.34
18	38	56	56	56	0.30
19	39	56	57	57	0.30
20	40	58	58	57	0.33
21	41	58	58	58	0.29
22	42	59	59	59	0.70
23	43	60	59	59	0.33
24	44	60	60	60	0.30
25	45	60	60	60	0.48
26	46	61	61	61	0.44
27	47	61	61	61	0.42
28	48	62	62	62	0.31
29	49	62	62	62	0.28
30	50	62	63	63	0.64
31	51	63	63	63	0.52
32	52	64	64	64	0.36
33	53	64	64	65	0.34
34	54	64	65	65	0.36
35	55	65	65	66	0.95
36	56	66	66	67	0.63
37	57	67	67	67	1.23

38	58	68	68	68	0.62
39	59	69	69	69	1.12
40	60	69	69	69	0.96
41	61	70	70	70	0.47
42	62	70	71	71	0.44
43	63	71	72	71	1.70
44	64	72	72	72	1.84
45	65	73	73	73	1.37
46	66	75	74	74	3.16
47	67	78	75	75	1.00
48	68	78	77	76	1.00
49	69	78	78	77	0.61
50	70	78	79	79	0.61
51	71	78	80	80	0.61
52	72	78	81	81	0.61
53	73	78	83	82	1.41
54	74	82	84	84	1.41
55	75	86	85	85	0.01
56	76	87	86	86	0.01
57	77	88	87	87	0.01
58	78	89	89	89	0.01
59	79	90	90	90	0.01

*The scores in the first column conform to the scoring rules of the CES-D Children instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 6: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From CES-D Children to PROMIS Pediatric Depressive Symptoms -Table 4 is recommended

CES-D Children Actual Score*	CES-D Children Raw Score*	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	20	30	28	26
1	21	32	33	32
2	22	35	35	35
3	23	37	37	37
4	24	39	39	39
5	25	40	40	41
6	26	42	42	43
7	27	44	44	44
8	28	46	46	46
9	29	47	47	47
10	30	49	49	49
11	31	50	50	50
12	32	52	51	51
13	33	52	52	52
14	34	53	53	53
15	35	54	54	54
16	36	54	55	54
17	37	55	56	55
18	38	56	56	56
19	39	57	57	57
20	40	58	58	57
21	41	58	58	58
22	42	59	59	59
23	43	60	60	59
24	44	60	60	60
25	45	61	60	60
26	46	61	61	61
27	47	61	62	62
28	48	62	62	62
29	49	62	62	62
30	50	63	63	63
31	51	63	63	64
32	52	64	64	64
33	53	64	64	65
34	54	64	65	65

35	55	65	66	66
36	56	66	66	66
37	57	67	67	67
38	58	68	68	68
39	59	69	68	68
40	60	69	69	69
41	61	70	70	70
42	62	71	71	70
43	63	71	71	71
44	64	71	72	72
45	65	73	73	73
46	66	75	74	73
47	67	77	75	74
48	68	78	75	75
49	69	78	76	76
50	70	78	77	76
51	71	78	78	77
52	72	78	78	78
53	73	79	79	79
54	74	80	80	80
55	75	82	81	81
56	76	82	82	82
57	77	83	83	83
58	78	84	84	84
59	79	85	85	85

*The scores in the first column conform to the scoring rules of the CES-D Children instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 7: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QoL v1.0 Pediatric Depression Full Item Bank and PROMIS Pediatric Depressive Symptoms (Neuro-QoL Wave 1 Study) - RECOMMENDED

Neuro-QoL Peds Depression T-Score	Neuro-QoL Peds Depression Raw Score	PROMIS Peds Depression T-Score	SE
32.0	17	31.3	5.5
36.2	18	35.4	4.6
38.3	19	37.7	4.3
40.5	20	40.0	3.8
42.1	21	41.8	3.4
43.5	22	43.4	3.2
44.6	23	44.7	2.9
45.6	24	45.9	2.8
46.4	25	47.0	2.6
47.2	26	47.9	2.5
47.8	27	48.8	2.4
48.5	28	49.7	2.3
49.1	29	50.4	2.2
49.7	30	51.2	2.1
50.2	31	51.9	2.1
50.7	32	52.5	2.0
51.2	33	53.2	2.0
51.7	34	53.8	2.0
52.2	35	54.4	2.0
52.7	36	55.0	1.9
53.2	37	55.6	1.9
53.7	38	56.1	1.9
54.1	39	56.7	1.9
54.6	40	57.2	1.9
55.1	41	57.8	1.9
55.6	42	58.3	1.9
56.0	43	58.8	1.9
56.5	44	59.4	1.9
57.0	45	59.9	1.9
57.5	46	60.5	1.9
58.0	47	61.0	1.9
58.5	48	61.5	1.9

59.0	49	62.0	1.9
59.5	50	62.6	1.9
60.0	51	63.1	1.9
60.5	52	63.6	1.9
61.0	53	64.2	1.9
61.4	54	64.7	1.9
61.9	55	65.2	1.9
62.4	56	65.7	1.9
62.9	57	66.3	1.9
63.4	58	66.8	1.9
63.8	59	67.3	1.9
64.3	60	67.9	1.9
64.8	61	68.4	1.9
65.2	62	68.9	1.9
65.7	63	69.4	1.9
66.1	64	70.0	1.9
66.6	65	70.5	1.9
67.0	66	71.0	1.9
67.5	67	71.6	1.9
67.9	68	72.1	1.9
68.4	69	72.6	1.9
68.8	70	73.2	1.9
69.3	71	73.8	1.9
69.8	72	74.3	1.9
70.3	73	74.9	1.9
70.8	74	75.5	1.9
71.3	75	76.2	2.0
71.9	76	76.8	2.0
72.4	77	77.5	2.0
73.1	78	78.2	2.1
73.7	79	79.0	2.2
74.5	80	79.9	2.3
75.3	81	80.8	2.4
76.3	82	81.9	2.5
77.5	83	83.0	2.6
79.1	84	84.3	2.6
81.6	85	85.7	2.5

Appendix Table 8: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL v1.0 Pediatric Depression Full Item Bank to PROMIS Pediatric Depressive Symptoms – Table 7 is recommended

Neuro-QoL Ped Depression Raw 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)
17	35	30	29	0.09
18	35	35	35	0.09
19	39	38	38	0.25
20	40	39	39	0.17
21	40	40	40	0.14
22	43	42	42	0.28
23	43	43	43	0.24
24	44	44	44	0.18
25	47	46	46	0.25
26	47	47	47	0.20
27	48	48	48	0.23
28	48	48	48	0.20
29	50	49	49	0.34
30	50	51	51	0.33
31	52	52	52	0.18
32	52	52	52	0.19
33	53	53	53	0.23
34	53	53	53	0.24
35	54	54	54	0.30
36	54	54	54	0.26
37	55	55	55	0.24
38	55	55	55	0.21
39	55	56	56	0.19
40	56	56	57	0.49
41	58	58	57	0.34
42	58	58	58	0.33
43	59	59	59	0.34
44	59	59	59	0.28
45	60	59	59	0.25
46	60	60	60	0.23
47	60	60	60	0.23
48	60	60	60	0.23
49	61	61	61	0.20
50	61	61	61	0.19
51	62	62	62	0.34
52	62	63	63	0.31
53	63	63	63	0.50
54	64	64	64	0.29
55	64	64	64	0.26

56	65	65	65	0.36
57	65	65	65	0.30
58	65	65	65	0.31
59	66	66	66	0.31
60	66	66	66	0.25
61	66	67	67	0.22
62	68	67	67	0.61
63	68	68	68	0.66
64	69	69	69	1.41
65	70	69	69	0.34
66	70	70	70	0.32
67	70	70	71	0.36
68	71	71	71	0.53
69	72	72	73	1.41
70	75	74	74	1.00
71	76	75	75	0.01
72	76	75	75	0.01
73	76	75	75	0.01
74	76	76	75	0.01
75	76	76	76	0.35
76	76	76	76	0.35
77	76	76	76	0.35
78	76	76	76	0.35
79	76	76	76	0.35
80	76	77	77	0.35
81	78	78	78	0.02
82	82	81	81	0.01
83	82	84	84	0.01
84	82	86	86	0.35
85	82	89	89	0.35

Appendix Table 9: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL v1.0 Pediatric Depression Full Item Bank to PROMIS Pediatric Depressive Symptoms – Table 7 is recommended

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

55	64	64	64
56	65	65	65
57	65	65	65
58	66	66	66
59	66	66	66
60	66	66	67
61	67	67	67
62	67	67	68
63	68	68	68
64	68	68	69
65	69	69	70
66	70	70	70
67	70	71	71
68	71	72	72
69	72	73	73
70	75	74	74
71	76	75	74
72	76	75	75
73	76	75	75
74	76	76	75
75	76	76	76
76	76	76	76
77	76	76	76
78	76	77	77
79	76	77	77
80	77	77	78
81	78	78	78
82	79	79	79
83	81	80	80
84	82	81	81
85	83	83	83

Appendix Table 10: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for SMFQ and PROMIS Pediatric Depressive Symptoms (PROsetta Stone Wave 2 Study) – RECOMMENDED

SMFQ Actual Score*	SMFQ Raw Score*	PROMIS Peds Depression T-Score	SE
0	13	37.8	6.5
1	14	42.8	5.5
2	15	46.3	4.9
3	16	49.3	4.3
4	17	51.7	3.9
5	18	53.8	3.5
6	19	55.6	3.2
7	20	57.2	3.1
8	21	58.7	2.9
9	22	60.1	2.8
10	23	61.4	2.8
11	24	62.7	2.7
12	25	64.0	2.7
13	26	65.2	2.7
14	27	66.4	2.7
15	28	67.6	2.7
16	29	68.8	2.7
17	30	70.0	2.7
18	31	71.3	2.7
19	32	72.6	2.7
20	33	73.9	2.8
21	34	75.3	2.9
22	35	76.7	3.0
23	36	78.3	3.1
24	37	80.1	3.3
25	38	81.9	3.4
26	39	83.8	3.3

*The scores in the first column conform to the scoring rules of the SMFQ instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 11: Direct (Raw to Scale) Equipercentile Crosswalk Table – From SMFQ to PROMIS Pediatric Depressive Symptoms – Table 10 is recommended

SMFQ Actual Score*	SMFQ Raw 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	13	35	35	35	0.31
1	14	41	41	41	0.24
2	15	45	45	45	0.46
3	16	49	49	49	0.62
4	17	52	52	52	0.24
5	18	54	54	54	0.45
6	19	56	56	56	0.37
7	20	57	57	57	0.77
8	21	58	58	58	0.29
9	22	60	59	59	0.40
10	23	61	61	61	0.56
11	24	62	62	62	0.35
12	25	62	63	63	0.31
13	26	64	64	64	0.51
14	27	65	65	65	1.05
15	28	66	66	66	0.68
16	29	68	67	68	0.76
17	30	68	69	69	0.70
18	31	70	70	70	0.50
19	32	70	71	71	2.74
20	33	76	74	74	2.83
21	34	78	77	77	1.41
22	35	79	79	79	2.00
23	36	82	82	82	1.41
24	37	88	84	84	1.41
25	38	89	87	87	1.41
26	39	90	89	89	1.41

*The scores in the first column conform to the scoring rules of the SMFQ instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 12: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From SMFQ to PROMIS Pediatric Depressive Symptoms – Table 10 is recommended

SMFQ Actual Score*	SMFQ Raw Score*	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
0	13	35	35	35
1	14	41	41	41
2	15	45	45	45
3	16	49	49	49
4	17	52	52	51
5	18	54	54	53
6	19	56	56	55
7	20	57	57	57
8	21	58	58	58
9	22	60	60	59
10	23	61	61	61
11	24	62	62	62
12	25	62	63	63
13	26	64	64	64
14	27	65	65	65
15	28	66	66	66
16	29	67	67	68
17	30	69	68	69
18	31	70	70	70
19	32	71	71	71
20	33	76	73	73
21	34	78	74	74
22	35	79	76	76
23	36	82	78	78
24	37	83	80	80
25	38	84	82	82
26	39	85	84	84

*The scores in the first column conform to the scoring rules of the SMFQ instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 13: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Pediatric FACIT Fatigue and PROMIS Pediatric Fatigue (PROsetta Stone Wave 2 Study) – RECOMMENDED

Peds FACIT Fatigue Actual Score*	Peds FACIT Fatigue Raw Score*	PROMIS Peds FatigueT-Score	SE
52	13	29.6	5.7
51	14	33.5	5.1
50	15	36.7	4.6
49	16	39.3	4.2
48	17	41.5	3.9
47	18	43.4	3.6
46	19	45.1	3.4
45	20	46.5	3.2
44	21	47.9	3.1
43	22	49.1	3.0
42	23	50.3	2.9
41	24	51.3	2.8
40	25	52.4	2.7
39	26	53.3	2.7
38	27	54.2	2.6
37	28	55.1	2.6
36	29	56.0	2.6
35	30	56.8	2.6
34	31	57.7	2.5
33	32	58.5	2.5
32	33	59.3	2.5
31	34	60.0	2.5
30	35	60.8	2.5
29	36	61.6	2.5
28	37	62.3	2.5
27	38	63.1	2.5
26	39	63.8	2.5
25	40	64.6	2.5
24	41	65.3	2.5
23	42	66.1	2.5
22	43	66.8	2.5
21	44	67.6	2.5
20	45	68.3	2.5
19	46	69.1	2.5
18	47	69.8	2.5
17	48	70.6	2.5
16	49	71.4	2.5
15	50	72.1	2.5
14	51	72.9	2.5
13	52	73.7	2.5
12	53	74.5	2.6

11	54	75.4	2.6
10	55	76.3	2.6
9	56	77.2	2.7
8	57	78.1	2.7
7	58	79.1	2.8
6	59	80.1	2.9
5	60	81.2	2.9
4	61	82.4	3.0
3	62	83.5	2.9
2	63	84.7	2.8
1	64	85.7	2.6
0	65	86.6	2.3

*The scores in the first column conform to the scoring rules of the Peds FACIT Fatigue instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 14: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Pediatric FACIT Fatigue to PROMIS Pediatric Fatigue – Table 13 is recommended

Peds FACIT Fatigue Actual Score*	Peds FACIT Fatigue Raw Score*	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
52	13	28	22	22	0.09
51	14	28	28	29	0.09
50	15	32	32	32	0.39
49	16	36	36	35	0.67
48	17	39	39	39	0.58
47	18	42	41	41	0.42
46	19	43	43	44	0.86
45	20	45	45	46	0.72
44	21	47	47	47	0.80
43	22	50	49	49	0.37
42	23	50	50	50	0.36
41	24	51	51	51	0.78
40	25	52	52	52	0.35
39	26	53	53	53	0.58
38	27	54	54	54	1.37
37	28	56	55	55	0.37
36	29	56	56	56	0.33
35	30	58	57	57	0.96
34	31	58	58	58	0.88
33	32	59	59	59	0.42
32	33	59	59	59	0.40
31	34	60	60	60	0.82
30	35	60	60	60	0.24
29	36	61	61	61	0.24
28	37	61	61	62	0.23
27	38	61	62	62	0.22
26	39	63	63	63	0.68
25	40	64	64	64	0.45
24	41	66	65	65	0.61
23	42	66	66	66	0.60
22	43	67	67	67	0.53
21	44	67	67	67	0.50
20	45	68	68	68	0.98

19	46	68	69	69	0.88
18	47	70	70	70	4.24
17	48	72	71	71	0.70
16	49	72	72	72	0.57
15	50	72	72	72	0.49
14	51	72	73	73	0.49
13	52	73	74	74	2.45
12	53	75	75	75	1.37
11	54	77	77	76	2.83
10	55	78	78	78	2.45
9	56	81	80	79	2.00
8	57	86	81	80	1.06
7	58	86	82	81	1.06
6	59	86	83	82	1.06
5	60	86	85	84	1.06
4	61	86	86	85	1.06
3	62	87	87	86	1.06
2	63	88	88	87	1.06
1	64	89	89	89	1.06
0	65	90	90	90	1.06

*The scores in the first column conform to the scoring rules of the Peds FACIT Fatigue instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 15: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – Pediatric FACIT Fatigue to PROMIS Pediatric Fatigue – Table 13 is recommended

Peds FACIT Fatigue Actual Score*	Peds FACIT Fatigue Raw Score*	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
52	13	27	26	23
51	14	30	30	30
50	15	33	33	33
49	16	36	36	36
48	17	39	38	39
47	18	41	41	41
46	19	43	43	43
45	20	45	45	45
44	21	47	48	47
43	22	50	49	49
42	23	50	50	50
41	24	51	51	51
40	25	52	52	52
39	26	53	53	53
38	27	54	54	54
37	28	55	55	55
36	29	56	56	56
35	30	58	57	57
34	31	58	58	58
33	32	58	58	58
32	33	59	59	59
31	34	60	60	60
30	35	60	60	60
29	36	61	61	61
28	37	61	61	62
27	38	62	62	62
26	39	63	63	63
25	40	64	64	64
24	41	66	65	65
23	42	66	66	66
22	43	67	67	66
21	44	67	68	67
20	45	68	68	68
19	46	69	69	69
18	47	70	70	70
17	48	71	70	71
16	49	72	71	72
15	50	72	72	72
14	51	73	73	73
13	52	73	74	74

12	53	75	75	75
11	54	77	77	76
10	55	78	78	78
9	56	81	80	79
8	57	86	81	79
7	58	86	81	80
6	59	86	82	81
5	60	86	83	82
4	61	86	83	83
3	62	86	84	84
2	63	86	85	84
1	64	86	85	85
0	65	86	86	86

*The scores in the first column conform to the scoring rules of the Peds FACIT Fatigue instrument. The scores in the second column correspond to the scores used in the full PROsetta Stone report.

Appendix Table 16: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Pediatric Mobility to PROMIS Pediatric PF-Mobility Bank– Table 17 (Less Smoothing) is recommended

Neuro-QoL Ped Mobility Raw 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)
31	15	13	13	0.03
32	15	15	14	0.03
33	15	15	15	0.03
34	15	15	15	0.03
35	15	15	15	0.03
36	15	16	16	0.03
37	16	16	16	0.16
38	16	16	16	0.12
39	16	16	16	0.11
40	16	16	16	0.11
41	17	16	17	0.23
42	17	17	17	0.21
43	17	17	17	0.21
44	17	17	17	0.21
45	17	17	17	0.22
46	18	18	18	0.31
47	18	18	18	0.28
48	18	18	18	0.25
49	18	18	18	0.22
50	18	19	19	0.22
51	20	19	19	0.27
52	20	20	20	0.26
53	20	20	20	0.24
54	20	20	20	0.20
55	20	20	20	0.18
56	20	21	21	0.16
57	22	21	21	0.28
58	22	22	22	0.26
59	22	22	22	0.29
60	22	22	22	0.30
61	22	22	22	0.35
62	23	23	23	1.25
63	23	23	23	1.15
64	24	23	23	0.29
65	24	24	24	0.29
66	24	24	24	0.30
67	24	24	24	0.29

68	24	24	24	0.30
69	24	24	24	0.28
70	25	25	25	0.58
71	25	25	25	0.60
72	25	25	25	0.59
73	25	25	25	0.63
74	26	25	25	0.57
75	26	26	26	0.54
76	26	26	26	0.54
77	26	26	26	0.55
78	26	26	26	0.57
79	26	26	26	0.59
80	26	26	26	0.59
81	26	27	27	0.58
82	26	27	27	1.33
83	27	27	27	1.33
84	28	28	28	1.15
85	28	28	28	1.12
86	28	28	28	0.71
87	29	29	28	0.71
88	29	29	29	0.61
89	29	29	29	0.61
90	29	29	29	0.61
91	29	29	29	0.61
92	29	29	29	0.61
93	29	29	29	0.61
94	29	29	29	0.61
95	29	29	29	0.61
96	29	30	30	0.57
97	29	30	30	0.57
98	29	30	30	0.57
99	30	30	30	1.22
100	30	30	31	0.94
101	30	31	32	0.94
102	31	33	34	0.79
103	42	41	37	0.35
104	42	42	43	0.41
105	42	43	44	0.35
106	42	44	45	0.35
107	42	45	46	0.35
108	42	47	48	0.38
109	42	49	50	0.38
110	42	52	52	0.38
111	42	54	54	0.38
112	42	56	56	0.38
113	54	58	58	1.41

114	54	60	60	1.41
115	54	62	62	1.41
116	54	64	65	1.41
117	79	66	67	0.02
118	80	68	69	0.02
119	81	71	71	0.02
120	82	73	73	0.02
121	83	75	75	0.02
122	84	77	77	0.02
123	85	79	79	0.02
124	86	81	81	0.02
125	87	83	83	0.02
126	88	85	85	0.02
127	89	87	87	0.02
128	90	89	89	0.02

Appendix Table 17: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Pediatric Mobility to PROMIS Pediatric PF-Mobility Bank– Less Smoothing (3rd column) is RECOMMENDED

Neuro-QoL Ped Mobility Raw Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
31	14	14	14
32	15	15	14
33	15	15	15
34	15	15	15
35	15	15	15
36	15	15	15
37	16	16	16
38	16	16	16
39	16	16	16
40	16	16	16
41	17	16	16
42	17	17	17
43	17	17	17
44	17	17	17
45	18	18	18
46	18	18	18
47	18	18	18
48	18	18	18
49	19	19	19
50	19	19	19
51	19	19	19
52	20	20	20
53	20	20	20
54	20	20	20
55	20	20	20
56	21	21	21
57	21	21	21
58	21	21	21
59	22	22	22
60	22	22	22
61	22	22	22
62	22	22	22
63	23	23	23
64	23	23	23
65	23	23	23
66	24	24	23
67	24	24	24
68	24	24	24
69	24	24	24

70	24	24	24
71	24	24	24
72	25	25	25
73	25	25	25
74	25	25	25
75	25	25	25
76	25	26	25
77	26	26	26
78	26	26	26
79	26	26	26
80	26	26	26
81	26	26	26
82	26	26	27
83	27	27	27
84	27	27	27
85	28	28	27
86	28	28	28
87	29	28	28
88	29	28	28
89	29	28	28
90	29	28	28
91	29	29	28
92	29	29	29
93	29	29	29
94	29	29	29
95	29	29	29
96	29	29	30
97	29	29	30
98	29	30	31
99	30	30	31
100	30	31	32
101	31	32	33
102	32	34	35
103	42	41	36
104	42	42	42
105	42	42	42
106	42	42	42
107	42	42	42
108	42	42	43
109	42	42	43
110	42	42	43
111	42	43	44
112	42	43	44
113	42	43	44
114	42	44	45
115	42	44	46
116	54	45	46
117	62	46	47

118	62	46	48
119	62	47	49
120	62	48	50
121	62	50	51
122	62	51	52
123	62	52	53
124	62	54	55
125	62	55	56
126	62	57	58
127	62	59	60
128	62	61	62

Appendix Table 18: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Neuro-QoL Pediatric Interaction with Peers and PROMIS Pediatric Peer Relationships (PROsetta Stone Wave 2 Study) - RECOMMENDED

Neuro-QoL Ped Interaction w/ Peers T-Score	Neuro-QoL Ped Interaction w/ Peers Raw Score	PROMIS T-score Pediatric Peer Relationships	SE
15.2	16	12.5	1.8
17.1	17	13.2	2.0
18.6	18	13.9	2.2
20.0	19	14.7	2.3
21.1	20	15.6	2.3
22.1	21	16.6	2.3
23.0	22	17.4	2.2
23.9	23	18.2	2.2
24.6	24	19.0	2.1
25.3	25	19.8	2.0
26.0	26	20.5	2.0
26.7	27	21.1	2.0
27.3	28	21.8	1.9
27.9	29	22.4	1.9
28.4	30	23.0	1.9
29.0	31	23.6	1.9
29.5	32	24.2	1.9
30.1	33	24.8	1.9
30.6	34	25.3	1.9
31.1	35	25.9	1.9
31.6	36	26.5	1.9
32.2	37	27.1	1.9
32.7	38	27.6	1.9
33.2	39	28.2	1.9
33.8	40	28.8	1.9
34.4	41	29.4	2.0
34.9	42	30.0	2.0
35.5	43	30.6	2.0
36.1	44	31.2	2.0
36.7	45	31.8	2.0
37.3	46	32.4	2.0
37.9	47	33.0	2.0

38.5	48	33.6	2.0
39.1	49	34.3	2.0
39.7	50	34.9	2.0
40.3	51	35.5	2.0
40.9	52	36.1	2.0
41.5	53	36.8	2.0
42.1	54	37.4	2.0
42.7	55	38.0	2.0
43.3	56	38.7	2.0
43.9	57	39.3	2.0
44.5	58	40.0	2.0
45.1	59	40.6	2.0
45.7	60	41.3	2.0
46.2	61	41.9	2.0
46.8	62	42.6	2.0
47.4	63	43.3	2.0
48.0	64	43.9	2.0
48.6	65	44.6	2.0
49.2	66	45.3	2.0
49.8	67	46.0	2.0
50.5	68	46.8	2.1
51.1	69	47.5	2.1
51.8	70	48.3	2.1
52.6	71	49.1	2.2
53.3	72	50.0	2.3
54.1	73	50.9	2.4
55.0	74	51.9	2.5
56.0	75	53.1	2.7
57.2	76	54.4	3.0
58.6	77	56.0	3.3
60.3	78	57.9	3.8
62.7	79	60.5	4.3
67.1	80	65.1	5.6

Appendix Table 19: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Pediatric Interaction with Peers to PROMIS Pediatric Peer Relationships – Table 18 is recommended

Neuro-QoL Ped Interaction w/ Peers Raw 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)
16	10	10	10	0.35
17	17	11	11	0.35
18	18	11	11	0.35
19	18	12	12	0.35
20	18	13	13	0.35
21	18	14	14	0.35
22	18	15	15	0.35
23	19	15	15	0.35
24	19	16	16	0.35
25	19	17	17	0.35
26	19	18	18	0.35
27	19	18	19	0.35
28	19	19	19	0.35
29	19	20	20	0.35
30	22	21	21	0.71
31	22	21	22	0.71
32	22	22	22	0.79
33	22	23	23	2.00
34	22	23	24	2.00
35	23	24	24	2.00
36	24	25	25	1.22
37	26	26	26	0.78
38	26	27	27	0.83
39	27	27	27	1.90
40	28	28	28	0.74
41	30	29	29	0.89
42	30	30	30	0.85
43	30	31	31	0.89
44	31	32	31	0.91
45	32	32	32	0.78
46	33	33	33	0.58
47	34	34	33	0.76
48	35	34	34	0.20
49	35	35	35	0.21
50	35	36	35	0.21
51	36	36	36	0.59
52	37	37	37	0.45
53	37	37	37	0.45
54	38	38	38	0.43

55	39	39	39	0.48
56	39	39	39	0.47
57	40	40	40	0.60
58	41	41	40	0.42
59	41	41	41	0.42
60	42	42	42	0.49
61	42	42	42	0.45
62	43	43	43	0.37
63	44	44	44	0.32
64	44	44	44	0.33
65	45	45	45	0.30
66	45	46	46	0.30
67	46	46	46	0.27
68	46	47	47	0.27
69	48	48	48	0.29
70	49	49	49	0.40
71	50	50	50	0.47
72	50	51	51	0.38
73	52	51	52	0.36
74	52	52	53	0.35
75	54	54	54	0.35
76	54	55	55	0.31
77	56	56	57	0.36
78	58	58	59	0.27
79	61	61	61	0.39
80	65	65	64	0.22

Appendix Table 20: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Pediatric Interaction with Peers to PROMIS Pediatric Peer Relationships – Table 18 is recommended

Neuro-QoL Ped Interaction w/ Peers Raw Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
16	15	14	14
17	17	15	15
18	17	16	16
19	18	16	16
20	19	16	17
21	19	17	17
22	19	17	18
23	19	18	18
24	19	18	19
25	19	18	19
26	19	19	20
27	19	20	20
28	19	20	21
29	19	20	22
30	22	21	22
31	22	22	23
32	23	22	24
33	23	23	24
34	23	24	25
35	23	25	26
36	24	26	26
37	26	26	27
38	26	27	28
39	27	28	28
40	28	28	29
41	30	29	30
42	30	30	30
43	30	31	31
44	31	32	31
45	32	32	32
46	33	33	33
47	34	34	33
48	34	34	34
49	35	35	34
50	35	35	35
51	36	36	36
52	36	36	36
53	37	37	37
54	38	38	38

55	38	38	38
56	39	39	39
57	40	40	39
58	40	40	40
59	41	41	41
60	41	42	41
61	42	42	42
62	43	43	43
63	44	44	44
64	44	44	44
65	45	45	45
66	46	46	46
67	46	46	46
68	47	47	47
69	48	48	48
70	49	49	49
71	50	50	50
72	51	51	51
73	52	52	52
74	52	53	53
75	54	54	54
76	55	55	56
77	56	57	57
78	58	59	59
79	60	61	60
80	65	64	63

Appendix Table 21: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for Pediatric PCF and Neuro-QoL Pediatric Applied Cognition General (PROsetta Stone Wave 2 Study) - RECOMMENDED

Peds PCF Raw Score	Neuro-QoL Ped Cognitive Function T-Score	SE	Peds PCF Raw Score	Neuro-QoL Ped Cognitive Function T-Score	SE
30	13.9	2.4	91	39.9	1.4
31	15.0	2.6	92	40.1	1.4
32	16.2	2.7	93	40.4	1.4
33	17.3	2.6	94	40.7	1.4
34	18.4	2.6	95	41.0	1.4
35	19.3	2.4	96	41.3	1.4
36	20.2	2.3	97	41.5	1.4
37	20.9	2.2	98	41.8	1.4
38	21.7	2.1	99	42.1	1.4
39	22.3	2.0	100	42.4	1.4
40	22.9	1.9	101	42.7	1.4
41	23.5	1.9	102	43.0	1.4
42	24.0	1.8	103	43.3	1.4
43	24.5	1.8	104	43.6	1.4
44	25.0	1.7	105	43.9	1.4
45	25.5	1.7	106	44.2	1.4
46	25.9	1.7	107	44.4	1.4
47	26.3	1.6	108	44.7	1.4
48	26.7	1.6	109	45.0	1.4
49	27.1	1.6	110	45.4	1.4
50	27.5	1.6	111	45.7	1.4
51	27.9	1.5	112	46.0	1.4
52	28.3	1.5	113	46.3	1.4
53	28.6	1.5	114	46.6	1.4
54	29.0	1.5	115	46.9	1.4
55	29.3	1.5	116	47.2	1.4
56	29.7	1.5	117	47.5	1.4
57	30.0	1.5	118	47.9	1.4
58	30.3	1.4	119	48.2	1.4
59	30.6	1.4	120	48.5	1.4
60	30.9	1.4	121	48.9	1.5
61	31.3	1.4	122	49.2	1.5
62	31.6	1.4	123	49.6	1.5
63	31.9	1.4	124	49.9	1.5
64	32.2	1.4	125	50.3	1.5
65	32.5	1.4	126	50.6	1.5
66	32.8	1.4	127	51.0	1.5
67	33.1	1.4	128	51.4	1.5

68	33.4	1.4	129	51.7	1.5
69	33.7	1.4	130	52.1	1.5
70	33.9	1.4	131	52.5	1.5
71	34.2	1.4	132	52.9	1.5
72	34.5	1.4	133	53.3	1.5
73	34.8	1.4	134	53.8	1.6
74	35.1	1.4	135	54.2	1.6
75	35.4	1.4	136	54.7	1.6
76	35.7	1.4	137	55.2	1.7
77	35.9	1.4	138	55.7	1.7
78	36.2	1.4	139	56.2	1.7
79	36.5	1.4	140	56.8	1.8
80	36.8	1.4	141	57.4	1.9
81	37.1	1.4	142	58.0	2.0
82	37.3	1.4	143	58.7	2.1
83	37.6	1.4	144	59.6	2.2
84	37.9	1.4	145	60.5	2.4
85	38.2	1.4	146	61.6	2.6
86	38.5	1.4	147	62.8	2.9
87	38.7	1.4	148	64.4	3.3
88	39.0	1.4	149	66.6	3.7
89	39.3	1.4	150	70.3	4.9
90	39.6	1.4			

Appendix Table 22: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Pediatric PCF to Neuro-QoL Pediatric Cognitive Function – Table 21 is recommended

Peds PCF Raw Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)	Standard Error of Equating (SEE)
30	10	10	10	0.35
31	11	11	11	0.35
32	19	11	12	0.35
33	20	12	12	0.35
34	25	13	13	0.35
35	25	14	14	0.35
36	25	15	15	0.35
37	25	15	16	0.35
38	25	16	16	0.35
39	25	17	17	0.35
40	25	18	18	0.35
41	25	19	19	0.35
42	25	19	20	0.35
43	25	20	20	0.35
44	25	21	21	0.35
45	25	22	22	0.35
46	25	23	23	0.35
47	25	23	24	0.35
48	25	24	25	0.35
49	25	25	25	0.35
50	25	26	26	0.35
51	28	27	27	0.71
52	28	27	27	0.71
53	28	28	28	0.71
54	28	28	28	0.79
55	28	29	29	0.79
56	30	29	29	2.00
57	30	30	29	2.00
58	31	30	30	0.50
59	31	30	30	0.52
60	31	31	30	0.52
61	31	31	31	0.52
62	31	31	31	0.56
63	31	31	31	0.56
64	31	31	31	0.54
65	32	32	32	0.77
66	32	32	32	0.90
67	32	32	32	1.02
68	32	33	33	1.70

69	34	33	33	0.34
70	34	33	33	0.32
71	34	34	34	0.32
72	34	34	34	0.34
73	34	34	34	0.34
74	34	34	34	0.32
75	34	35	35	0.35
76	35	35	35	2.15
77	36	36	36	0.71
78	36	36	36	0.70
79	36	36	36	0.45
80	37	37	37	0.43
81	37	37	37	0.44
82	37	37	37	0.44
83	37	37	37	0.46
84	38	38	38	0.41
85	38	38	38	0.40
86	38	38	38	0.43
87	39	39	39	0.32
88	39	39	39	0.31
89	39	39	39	0.27
90	40	40	40	0.33
91	40	40	40	0.34
92	40	40	40	0.35
93	41	41	41	0.35
94	41	41	41	0.35
95	41	41	41	0.35
96	41	41	41	0.35
97	42	42	42	0.43
98	42	42	42	0.43
99	42	42	42	0.43
100	42	42	42	0.45
101	43	43	43	0.44
102	43	43	43	0.44
103	43	43	43	0.42
104	43	44	44	0.43
105	44	44	44	0.64
106	44	44	44	0.62
107	44	45	45	0.61
108	45	45	45	0.69
109	45	45	45	0.72
110	46	46	46	0.41
111	46	46	46	0.41
112	47	47	47	0.54
113	47	47	47	0.52
114	48	47	47	0.24
115	48	48	48	0.23
116	48	48	48	0.22

117	48	48	48	0.22
118	48	48	48	0.22
119	49	49	49	0.38
120	49	49	49	0.36
121	49	49	49	0.37
122	50	50	50	0.28
123	50	50	50	0.28
124	50	50	50	0.27
125	50	51	51	0.27
126	51	51	51	0.92
127	52	51	51	0.23
128	52	52	52	0.22
129	52	52	52	0.23
130	53	53	53	0.48
131	53	53	53	0.46
132	54	53	53	0.48
133	54	54	54	0.48
134	54	54	54	0.48
135	55	54	54	0.31
136	55	55	55	0.30
137	55	55	55	0.30
138	55	55	56	0.30
139	56	56	56	0.37
140	56	56	57	0.33
141	56	57	57	0.33
142	58	57	58	0.36
143	58	58	58	0.35
144	59	59	59	0.35
145	59	60	60	0.31
146	62	61	61	0.20
147	62	62	63	0.18
148	66	65	64	0.16
149	66	66	65	0.12
150	66	72	71	0.11

Appendix Table 23: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Pediatric PCF to Neuro-QoL Pediatric Cognitive Function – Table 21 is recommended

Peds PCF Raw Score	Equipercentile PROMIS Scaled Score Equivalents (No Smoothing)	Equipercentile Equivalents with Postsmoothing (Less Smoothing)	Equipercentile Equivalents with Postsmoothing (More Smoothing)
30	17	18	18
31	17	19	19
32	19	19	19
33	24	20	20
34	24	21	21
35	25	21	21
36	25	22	22
37	25	22	22
38	25	23	23
39	25	23	23
40	25	23	23
41	25	24	24
42	25	24	24
43	25	24	24
44	25	25	25
45	25	25	25
46	25	25	25
47	25	25	26
48	25	26	26
49	25	26	26
50	25	26	26
51	28	27	27
52	28	28	27
53	28	28	28
54	28	29	28
55	29	29	28
56	30	29	29
57	30	30	29
58	31	30	30
59	31	30	30
60	31	31	30
61	31	31	31
62	31	31	31
63	31	31	31
64	31	32	32
65	32	32	32
66	32	32	32
67	32	32	33

68	33	33	33
69	33	33	33
70	33	33	34
71	33	34	34
72	34	34	34
73	34	34	34
74	34	34	35
75	34	35	35
76	35	35	35
77	36	35	36
78	36	36	36
79	36	36	36
80	37	36	36
81	37	37	37
82	37	37	37
83	37	37	37
84	38	38	38
85	38	38	38
86	38	38	38
87	38	38	38
88	38	39	39
89	39	39	39
90	40	39	39
91	40	40	40
92	40	40	40
93	41	40	40
94	41	41	41
95	41	41	41
96	41	41	41
97	42	42	42
98	42	42	42
99	42	42	42
100	42	42	42
101	43	43	43
102	43	43	43
103	43	43	43
104	44	44	44
105	44	44	44
106	44	44	44
107	45	45	45
108	45	45	45
109	45	45	45
110	45	46	46
111	46	46	46
112	47	46	46
113	47	47	46
114	47	47	47
115	48	48	47
116	48	48	48

117	48	48	48
118	48	48	48
119	48	48	48
120	49	49	49
121	49	49	49
122	49	49	50
123	49	50	50
124	50	50	50
125	50	50	50
126	51	51	51
127	51	51	51
128	52	52	52
129	53	52	52
130	53	53	52
131	54	53	53
132	54	54	53
133	54	54	54
134	54	54	54
135	55	55	54
136	55	55	55
137	55	55	55
138	56	56	56
139	56	56	56
140	57	57	57
141	57	57	57
142	58	58	58
143	58	58	58
144	59	59	59
145	60	60	60
146	61	61	61
147	62	62	62
148	64	63	64
149	65	65	66
150	68	68	69