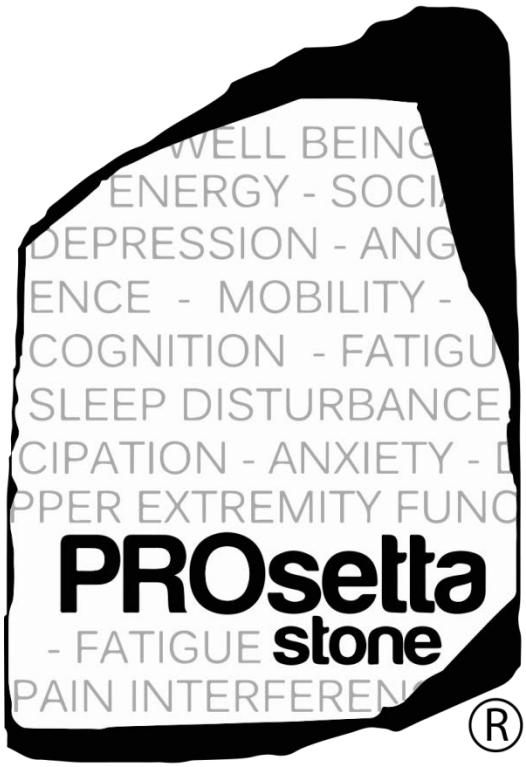

PROSETTA STONE® ANALYSIS REPORT

A ROSETTA STONE FOR PATIENT REPORTED OUTCOMES

PROMIS PEDIATRIC PHYSICAL FUNCTION-MOBILITY AND NEURO-QOL PEDIATRIC LOWER EXTREMITY - MOBILITY

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.

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-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.)

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

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

Developed in 2006 with the NIH Blueprint funding for Neuroscience Research, four domains of assessment central to neurological and behavioral function were created to measure cognition, sensation, motor functioning, and emotional health. The NIH Toolbox for Assessment of Neurological and Behavioral Function[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 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] www.nihtoolbox.org
[4] www.neuroqol.org

3. Legacy Instruments

Typically, we have linked widely accepted "legacy" measures that were part of the initial validation work for PROMIS or NIH Toolbox. In some cases, instruments were administered as part of the PROsetta Stone project for specific linking purposes. In this case, we have linked Neuro-QoL Pediatric Lower Extremity -Mobility to PROMIS Pediatric Physical Function - Mobility, where the former serves as the "legacy" instrument. 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).

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. Linking Assumptions

In Section 5 of this PROsetta Stone report, we present the results of a large number of linking studies using a combination of newly collected and secondary data sets. In most cases, we have applied all three linking methods described in sections 4.1 and 4.2. Our purpose is to provide the maximum amount of useful information. However, the suitability of these methods depends upon the meeting of various linking assumptions. These assumptions require that the two instruments to be linked measure the same construct, show a high correlation, and are relatively invariant in subpopulation differences (Dorans, 2007). The degree to which these assumptions are met varies across linking studies. Given that different researchers may interpret these requirements differently, we have taken a liberal approach for inclusion of linkages in this book. Nevertheless, we recommend that researchers diagnostically review the classical psychometrics and CFA results in light of these assumptions prior to any application of the cross-walk charts or legacy parameters to their own data. Having investigated a large number of possible links between PROMIS measures and legacy measures, we did apply a few minimal exclusion rules before linking. For example, we generally did not proceed with planned linking when the raw score correlation between two instruments was less than .70.

5. Linking Results

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

Table 5.1. Linking by Reference Instrument

Section	PROMIS Instrument	Instrument to Link	Study
5.6	PROMIS Pediatric Mobility	Neuro-QoL Pediatric Mobility	Neuro-QoL Wave 1

5.6 PROMIS Pediatric Physical Function - Mobility and Neuro-QoL Pediatric Lower Extremity -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 PF-Mobility item bank (eight item selection) and Neuro-QoL Pediatric Lower Extremity- Mobility scale (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.

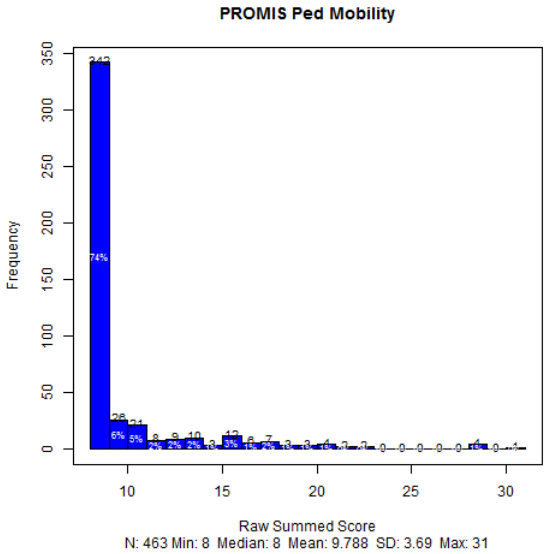

Figure 5.6.1: Raw Summed Score Distribution - PROMIS Pediatric Mobility

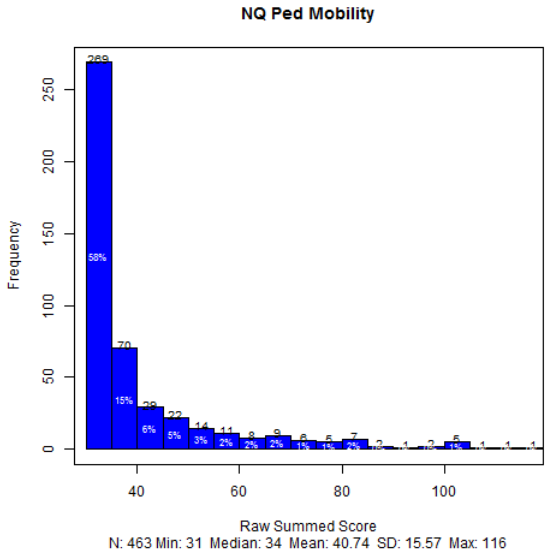

Figure 5.6.2: Raw Summed Score Distribution – Neuro-QoL Pediatric Mobility

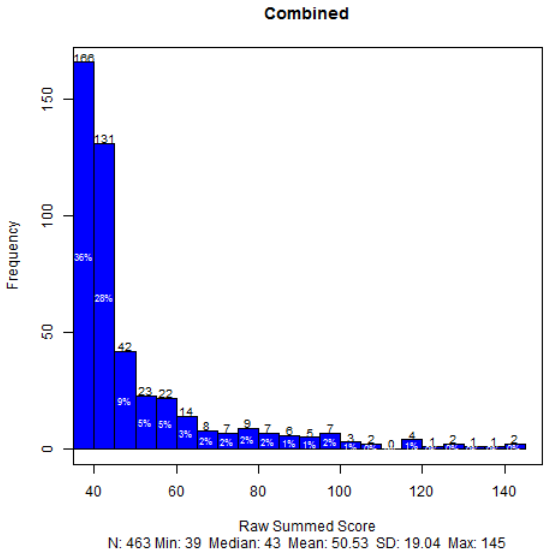

Figure 5.6.3: Raw Summed Score Distribution – Combined

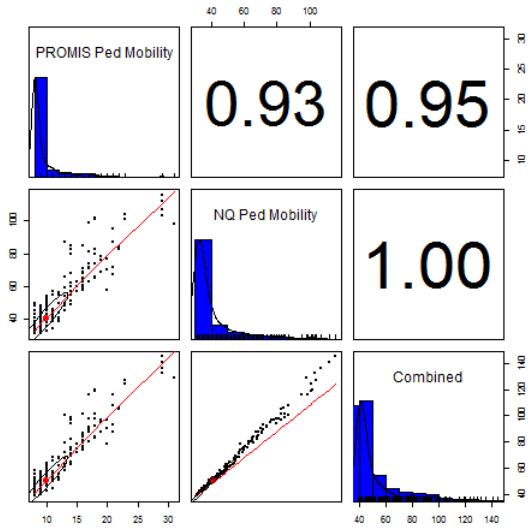

Figure 5.6.4: Scatter Plot Matrix of Raw Summed Scores

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 5.6.5 displays the cumulative distribution functions of the measures. Figure 5.6.6 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).

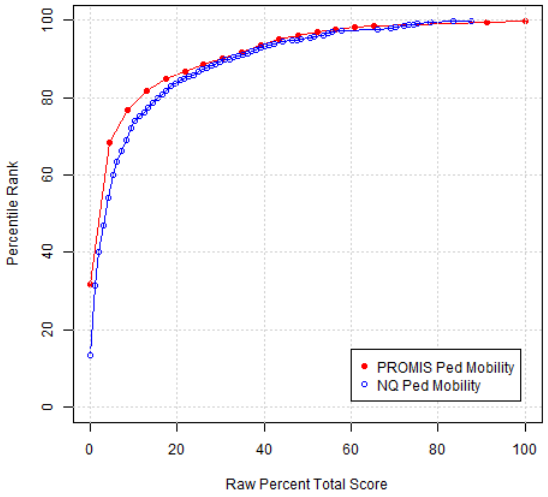

Figure 5.6.5: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

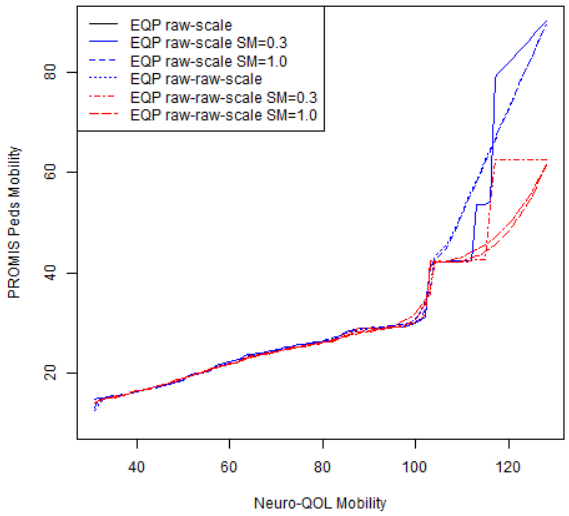

Figure 5.6.6: Equipercentile Linking Functions

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 5.6.3 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 5.6.4 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

6. Appendix Table 16: Direct (Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Pediatric LE-Mobility 31-Item Scale to PROMIS Pediatric PF-Mobility Item 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

7. Appendix Table 17: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table – From Neuro-QoL Pediatric LE-Mobility 31-Item Scale to PROMIS Pediatric PF-Mobility Item 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