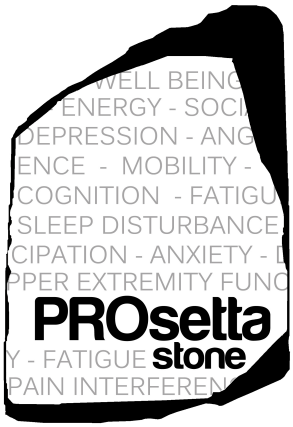

PRO Rosetta Stone Analysis

(PROSETTA Wave 1 Study)

Seung W. Choi, Tracy Podrabsky, Natalie McKinney,
and David Cella

Department of Medical Social Sciences
Feinberg School of Medicine
Northwestern University

March 21, 2012

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.

PROsetta Stone Analysis
PROMIS Depression and HADS Depression

Seung W. Choi, Tracy Podrabsky, Natalie McKinney, and David Cella
Department of Medical Social Sciences
Feinberg School of Medicine
Northwestern University

March 21, 2012

1 Introduction

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

only weaker forms of linking are meaningful, such as calibration, concordance, projection, or moderation.

2 The PRO Rosetta Stone Project

The primary aim of the PRO Rosetta Stone (PROsetta Stone^{TM}) project (1RC4CA157236-01, PI: David Cella) is to develop and apply methods to link the Patient-Reported Outcomes Measurement Information System (PROMIS) measures with other related "legacy" instruments to expand the range of PRO assessment options within a common, standardized metric. The project identifies and applies appropriate linking methods that allow scores on a range of assessment instruments to be expressed as standardized T-score metrics linked to the PROMIS. This preliminary report encompasses the first wave of 20 linking studies based on available PRO data from PROMIS (aka, PROMIS Wave I), Toolbox, and Neuro-QOL.

Add info here to describe PROSETTA Wave 1.

2.1 Patient-Reported Outcomes Measurement Information System (PROMIS)

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

All PROMIS banks have been calibrated according to Samejima's (1969) graded response model (based on large data collections including both general and clinical samples) and re-scaled (mean=50 and SD=10) using scale-setting subsamples matching the marginal distributions of gender, age, race,

[1] http://www.nihpromis.org
[2] http://www.nihroadmap.nih.gov

and education in the 2000 US census. The PROMIS Wave I calibration data included a small number of full-bank testing cases (approximately 1,000 per bank) from a general population taking one full bank and a larger number of block-administration cases (n= 14,000) from both general and clinical populations taking a collection of blocks representing all banks with 7 items each. The full-bank testing samples were randomly assigned to one of 7 different forms. Each form was composed of one or more PROMIS domains (with an exception of Physical Function where the bank was split over two forms) and one or more legacy measures of the same or related domains.

The PROMIS Wave I data collection design included a number of widely accepted "legacy" measures. The legacy measures used for validation evidence included Buss-Perry Aggression Questionnaire (BPAQ), Center for Epidemiological Studies Depression Scale (CES-D), Mood and Anxiety Symptom Questionnaire (MASQ), Functional Assessment of Chronic Illness Therapy (FACIT), Brief Pain Inventory (BPI), and SF-36. In addition to the pairs for validity (e.g., PROMIS Depression and CES-D), the PROMIS Wave I data allows for the potential for linking over a dozen pairs of measures/subscales. Furthermore, included within each of the PROMIS banks were items from many other existing measures. Depending on the nature and strength of relationship between the measures, various linking procedures can be used to allow for cross-walking of scores.

2.2 Beck Depression Inventory, second edition (BDI-I)

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

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

clarity.

2.3 Hospital Anxiety and Depresion Scale (HADS)

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

3 Linking Methods

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

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

In order to apply the linking methods consistently across different studies, linking/concordance relationships will be established based on the raw summed score metric of the measures. Furthermore, the direction of linking relationships to be established will be from legacy to PROMIS. That is, each raw summed score on a given legacy instrument will be mapped to a T-score of the corre-

sponding PROMIS instrument/bank. Finally, the raw summed score for each legacy instrument was constructed such that higher scores represent higher levels of the construct being measured. When the measures were scaled in the opposite direction, we reversed the direction of the legacy measure in order for the correlation between the measures to be positive and to facilitate concurrent calibration. As a result, some or all item response scores for some legacy instruments will need to be reverse-coded.

3.1 IRT Linking

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

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

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

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

accuracy of parameter estimation even when linking is not necessary (e.g., two samples are from the same population having the same or similar ability levels). Thus, conducting IRT linking would be worthwhile.

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

3.2 Equipercentile Linking

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

Under the current context, equipercentile crosswalk tables can be generated using two different approaches. First is a direct linking approach where each raw summed score on non-PROMIS measure is mapped directly to a PROMIS T-score. That is, raw summed scores on the non-PROMIS instrument and IRT scaled scores on the PROMIS (reference) instrument are linked directly, although raw summed scores and IRT scaled score have distinct properties (e.g.,

discrete vs. continuous). This approach might be appropriate when the reference instrument is either an item bank or composed of a large number of items and so various subsets (static or dynamic) are likely to be used but not the full bank in its entirety (e.g., PROMIS Physical Function bank with 124 items). Second is an indirect approach where raw summed scores on the non-PROMIS instrument are mapped to raw summed scores on the PROMIS instrument; and then the resulting raw summed score equivalents are mapped to corresponding scaled scores based on a raw-to-scale score conversion table. Because the raw summed score equivalents may take fractional values, such a conversion table will need to be interpolated using statistical procedures (e.g., cubic spline).

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

4 Linking PROMIS Depression and HADS Depression

In this section we provide a summary of the procedures employed to establish a crosswalk between two measures of Depression, namely the PROMIS Depression item bank (15 items) and HADS Depression (7 items). PROMIS Depression was scaled such that higher scores represent higher levels of Depression. We created raw summed scores for each of the measures separately and then for the combined. Summing of item scores assumes that all items have positive correlations with the total as examined in the section on Classical Item Analysis.

4.1 Raw Summed Score Distribution

The maximum possible raw summed scores were 75 for PROMIS Depression and 28 for HADS Depression. Figures 1 and 2 graphically display the raw summed score distributions of the two measures. Figure 3 shows the distribution for the combined. Figure 4 is a scatter plot matrix showing the relationship of each pair of raw summed scores. Pearson correlations are shown above the diagonal. The correlation between PROMIS Depression and HADS Depression was 0.71. The disattenuated (corrected for unreliabilies) correlation between PROMIS Depression and HADS Depression was 0.77. The correlations between the combined score and the measures were 0.98 and 0.82 for PROMIS Depression and HADS Depression, respectively.

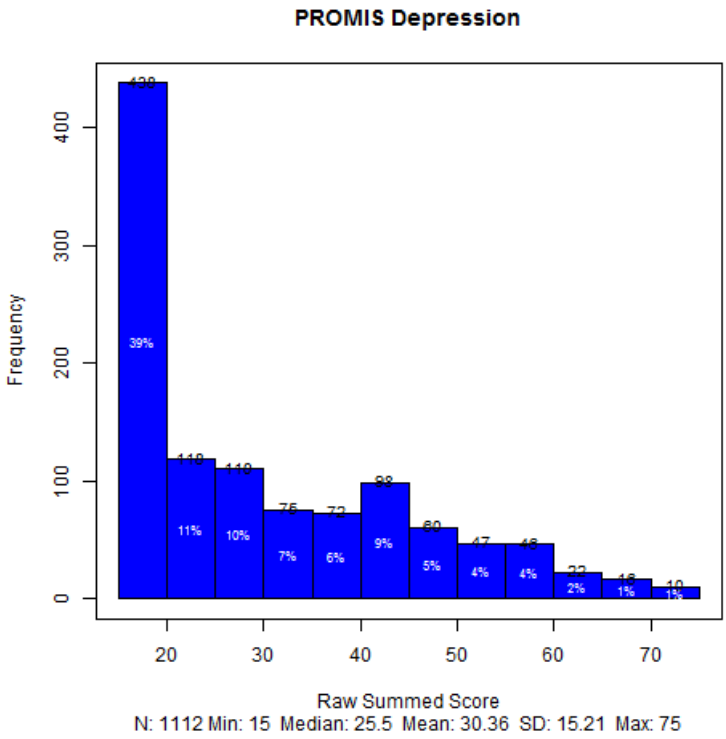

Figure 1: Raw Summed Score Distribution - PROMIS Depression

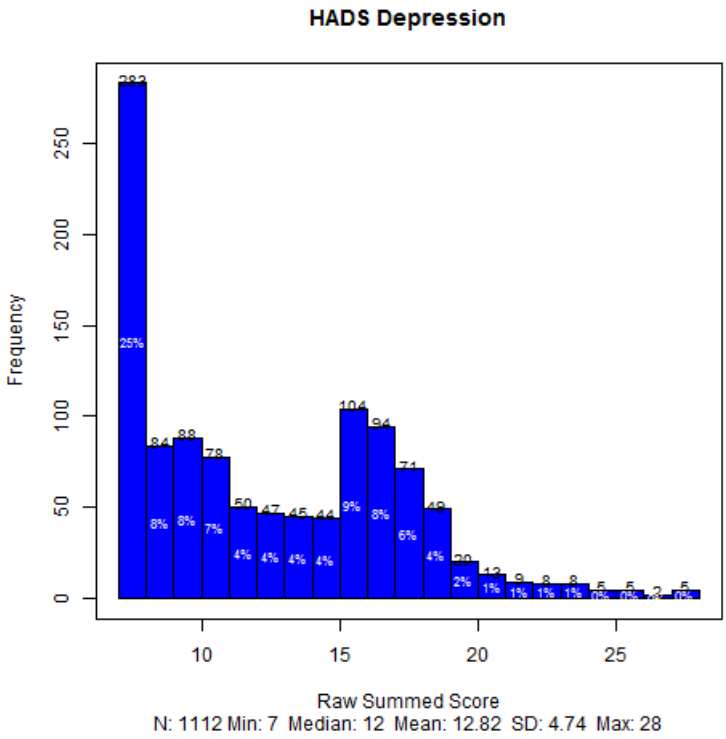

Figure 2: Raw Summed Score Distribution - HADS Depression

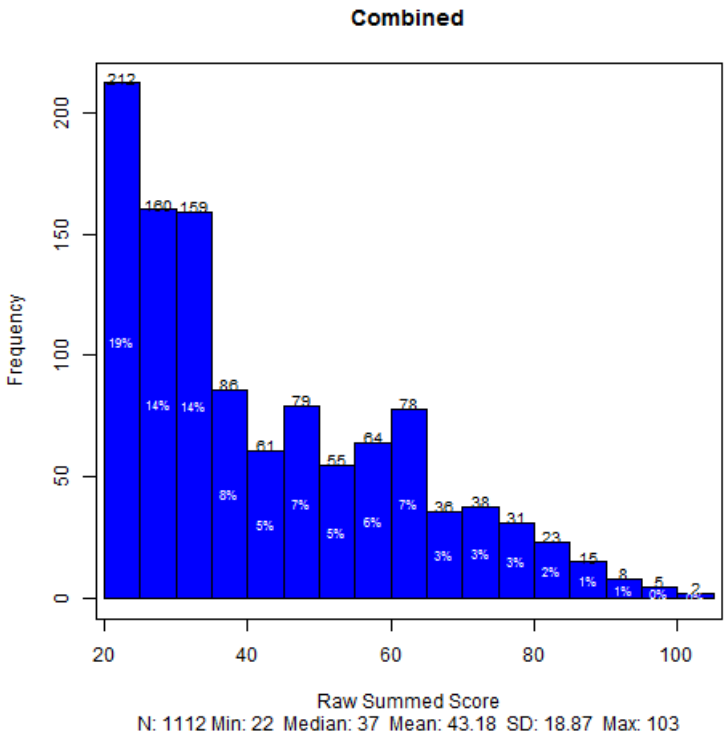

Figure 3: Raw Summed Score Distribution - Combined

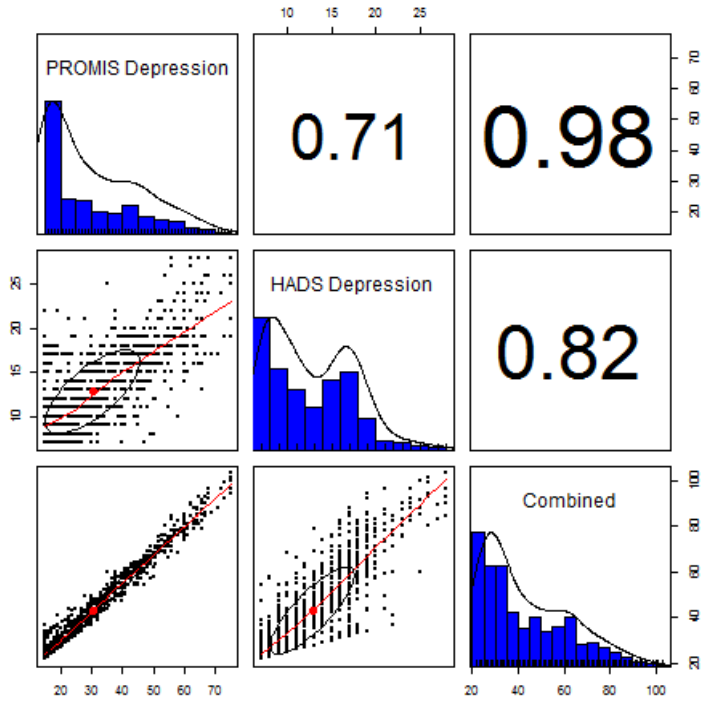

Figure 4: Scatter Plot Matrix of Raw Summed Scores

4.2 Classical Item Analysis

We conducted classical item analyses on the two measures separately and on the combined. Table 1 summarizes the results. For PROMIS Depression, Cronbach's alpha internal consistency reliability estimate was 0.98 and adjusted (corrected for overlap) item-total correlations ranged from 0.805 to 0.903. For HADS Depression, alpha was 0.859 and adjusted item-total correlations ranged from 0.572 to 0.7. For the 22 items, alpha was 0.972 and adjusted item-total correlations ranged from 0.451 to 0.893.

Table 1: Classical Item Analysis

	No. Items	Alpha	min.r	mean.r	max.r
PROMIS Depression	15	0.980	0.805	0.865	0.903
HADS Depression	7	0.859	0.572	0.630	0.700
Combined	22	0.972	0.451	0.767	0.893

4.3 Confirmatory Factor Analysis (CFA)

To assess the dimensionality of the measures, a categorical confirmatory factor analysis (CFA) was carried out using the WLSMV estimator of Mplus on a subset of cases without missing responses. A single factor model (based on polychoric correlations) was run on each of the two measures separately and on the combined. Table 2 summarizes the model fit statistics. For PROMIS Depression, the fit statistics were as follows: CFI = 0.994, TLI = 0.992, and RMSEA = 0.091. For HADS Depression, CFI = 0.925, TLI = 0.888, and RMSEA = 0.232. For the 22 items, CFI = 0.958, TLI = 0.954, and RMSEA = 0.155. The main interest of the current analysis is whether the combined measure is essentially unidimensional.

Table 2: CFA Fit Statistics

	No. Items	n	CFI	TLI	RMSEA
PROMIS Depression	15	1120	0.994	0.992	0.091
HADS Depression	7	1120	0.925	0.888	0.232
Combined	22	1120	0.958	0.954	0.155

4.4 Item Response Theory (IRT) Linking

We conducted concurrent calibration on the combined set of 22 items according to the graded response model. The calibration was run using `MULTILOG` and two different approaches as described previously (i.e., IRT linking vs. fixed-parameter calibration). For IRT linking, all 22 items were calibrated freely on

the conventional theta metric (mean=0, SD=1). Then the 15 PROMIS Depression items served as anchor items to transform the item parameter estimates for the HADS Depression items onto the PROMIS Depression metric. We used four IRT linking methods implemented in `plink` (Weeks, 2010): mean/mean, mean/sigma, Haebara, and Stocking-Lord. The first two methods are based on the mean and standard deviation of item parameter estimates, whereas the latter two are based on the item and test information curves. Table 3 shows the additive (A) and multiplicative (B) transformation constants derived from the four linking methods. For fixed-parameter calibration, the item parameters for the PROMIS Depression items were constrained to their final bank values, while the HADS Depression items were calibrated, under the constraints imposed by the anchor items.

Table 3: IRT Linking Constants

	A	B
Mean/Mean	1.252	0.330
Mean/Sigma	1.231	0.344
Haebara	1.216	0.365
Stocking-Lord	1.231	0.345

The item parameter estimates for the HADS Depression items were linked to the PROMIS Depression metric using the transformation constants shown in Table 3. The HADS Depression item parameter estimates from the fixed-parameter calibration are considered already on the PROMIS Depression metric. Based on the transformed and fixed-parameter estimates we derived test characteristic curves (TCC) for HADS Depression as shown in Figure 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 6 displays the differences on the vertical axis.

Table 4 shows the fixed-parameter calibration item parameter estimates for HADS Depression. The marginal reliability estimate for HADS Depression based on the item parameter estimates was 0.724. The marginal reliability estimates for PROMIS Depression and the combined set were 0.914 and 0.932, respectively. The slope parameter estimates for HADS Depression ranged from 0.888 to 1.92 with a mean of 1.38. The slope parameter estimates for PROMIS Depression ranged from 2.38 to 4.45 with a mean of 3.53. We also derived scale information functions based on the fixed-parameter calibration result. Figure 7 displays the scale information functions for PROMIS Depression, HADS Depression, and the combined set of 22. We then computed IRT scaled scores for the three measures based on the fixed-parameter calibration result. Figure 8 is a scatter plot matrix showing the relationships between the measures.

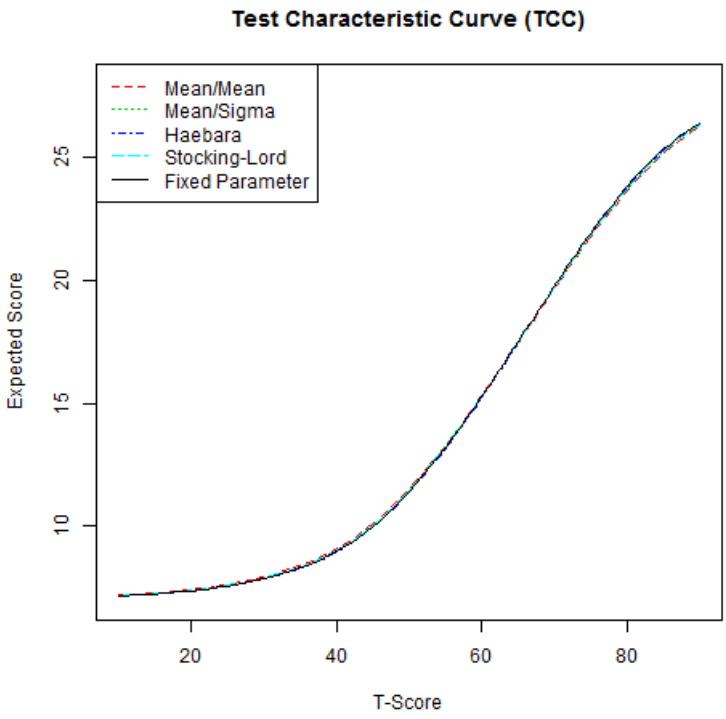

Figure 5: Test Characteristic Curves (TCC) from Different Linking Methods

Table 4: Fixed-Parameter Calibration Item Parameter Estimates for HADS Depression

a	cb1	cb2	cb3	NCAT
1.705	0.082	1.576	2.757	4
1.665	0.682	1.861	3.236	4
1.109	-0.220	1.542	2.835	4
0.888	-1.220	1.101	2.671	4
0.898	-0.101	1.267	2.877	4
1.919	0.189	1.349	2.574	4
1.486	0.553	1.838	2.891	4

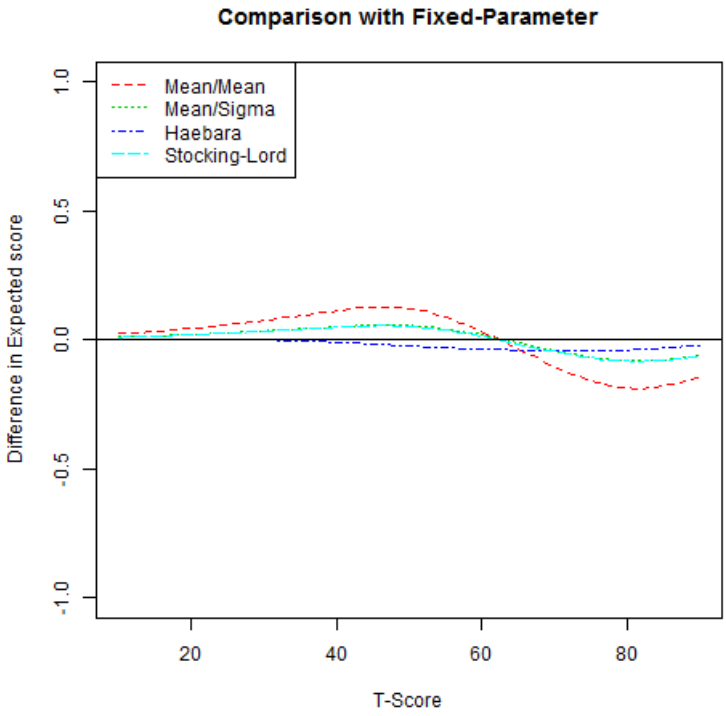

Figure 6: Difference in Test Characteristic Curves (TCC)

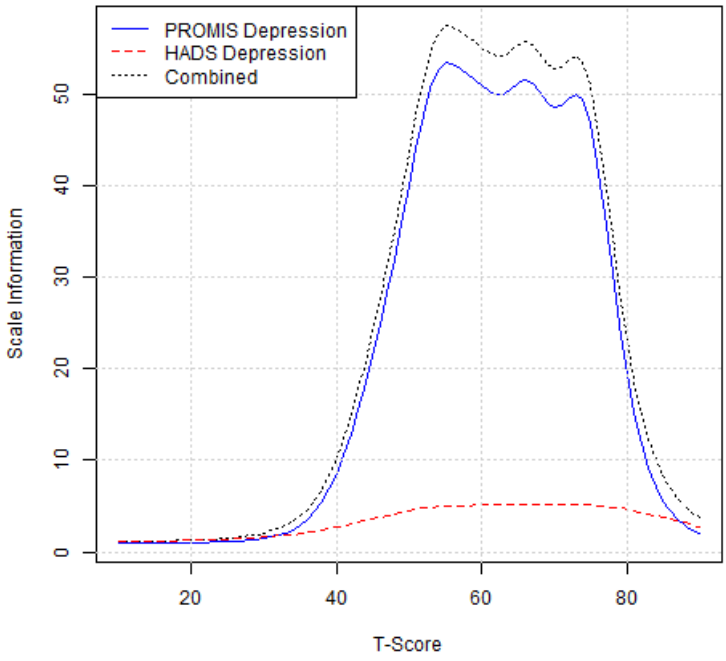

Figure 7: Comparison of Scale Information Functions

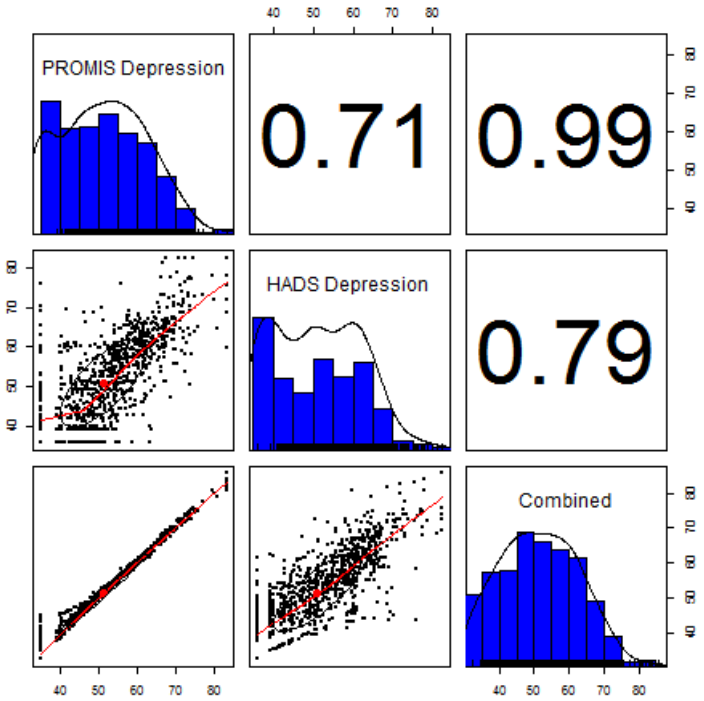

Figure 8: Comparison of IRT Scaled Scores

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 HADS Depression to a scaled score on PROMIS Depression can be useful. Based on the HADS Depression item parameters derived from the fixed-parameter calibration, we constructed a score conversion table. The conversion table displayed in Table 5 can be used to map simple raw summed scores from HADS Depression to T-score values linked to the PROMIS Depression metric. Each raw summed score point and corresponding PROMIS scaled score are presented along with the standard error associated with the scaled score. The raw summed score is constructed such that for each item, consecutive integers in base 1 are assigned to the ordered response categories.

Table 5: Raw Score to T-Score Conversion Table (IRT Fixed Parameter Calibration Linking) for HADS Depression

Raw	Tscore	SE
7	35.8	6.8
8	40.0	6.2
9	43.3	5.9
10	46.0	5.8
11	48.6	5.5
12	50.9	5.3
13	53.1	5.2
14	55.2	5.0
15	57.1	4.9
16	58.9	4.8
17	60.8	4.8
18	62.5	4.7
19	64.3	4.7
20	66.1	4.7
21	67.8	4.7
22	69.6	4.7
23	71.5	4.7
24	73.5	4.7
25	75.5	4.7
26	77.7	4.7
27	80.0	4.6
28	82.3	4.3

4.6 Equipercentile Linking

We mapped each raw summed score point on HADS Depression to a corresponding scaled score on PROMIS Depression by identifying scores on PROMIS Depression that have the same percentile ranks as scores on HADS 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 9 displays the cumulative distribution functions of the measures. Figure 10 shows the equipercentile linking functions based on raw summed scores, from HADS Depression to PROMIS Depression. When the number of raw summed score points differs substantially, the equipercentile linking functions could deviate from each other noticeably. The problem can be exacerbated when the sample size is small. Tables 6 and 7 show the equipercentile crosswalk tables. The result shown in Table 6 is based on the direct (raw summed score to scaled score) approach, whereas Table 7 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).

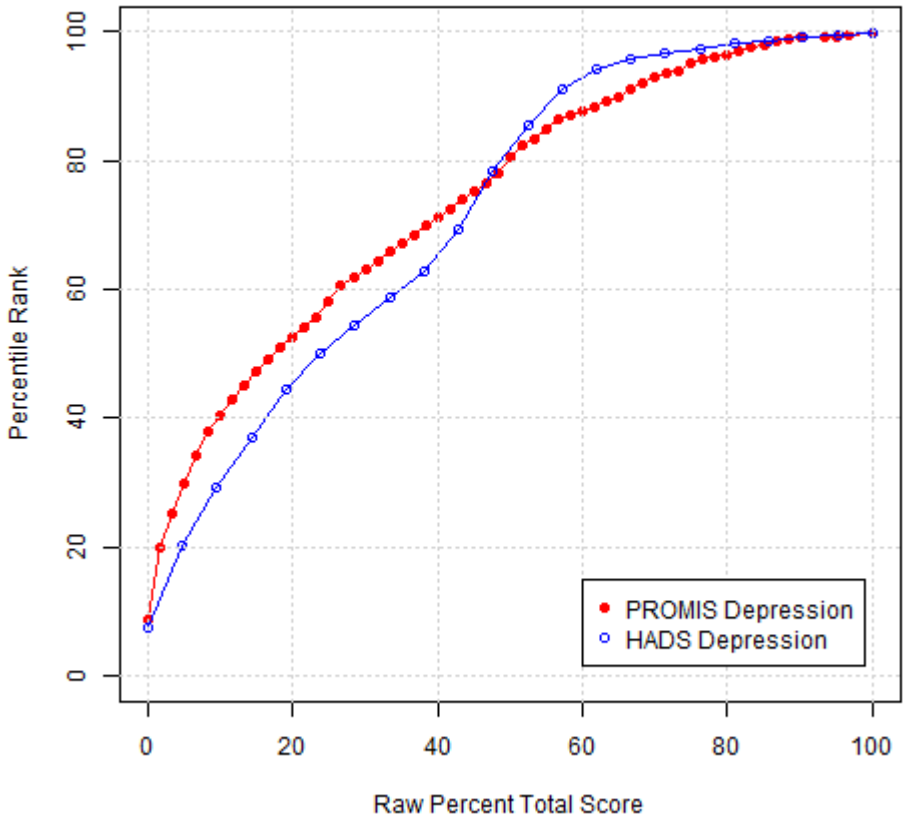

Figure 9: Comparison of Cumulative Distribution Functions based on Raw Summed Scores

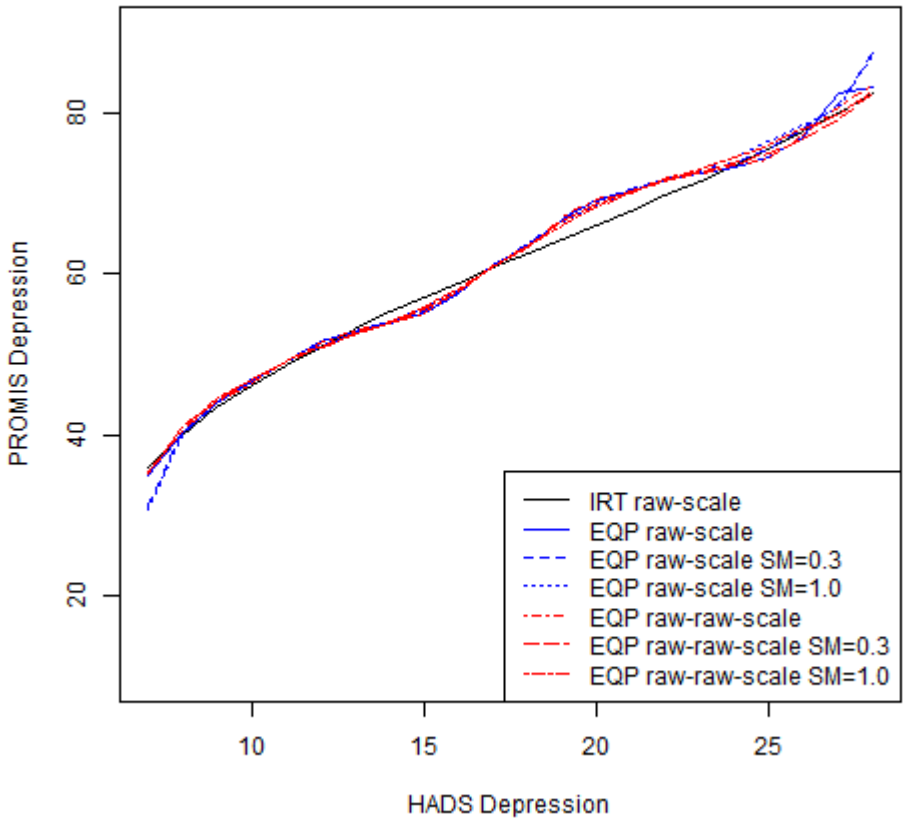

Figure 10: Equipercentile Linking Functions

Table 6: Direct (Raw to Scale) Equipercentile Crosswalk Table - From HADS Depression to PROMIS Depression

Score	Equi.EQ	Equi.SM.0.3	Equi.SM.1.0	SEE
7	35	31	31	0.26
8	40	40	40	0.26
9	44	44	44	0.37
10	47	47	47	0.61
11	49	49	49	0.32
12	52	51	51	0.44
13	53	53	52	0.45
14	54	54	54	0.30
15	55	55	56	0.52
16	58	58	58	0.63
17	61	61	61	0.86
18	63	64	64	0.40
19	67	67	66	0.44
20	69	69	69	0.72
21	70	70	70	0.35
22	72	72	72	0.55
23	72	73	73	0.62
24	73	74	75	0.50
25	74	75	76	0.89
26	77	78	78	4.90
27	82	81	81	0.60
28	83	87	87	0.46

Table 7: Indirect (Raw to Raw to Scale) Equipercentile Crosswalk Table - From HADS Depression to PROMIS Depression

Score	Equi.EQ	Equi.SM.0.3	Equi.SM.1.0
7	35	35	35
8	40	41	41
9	44	44	44
10	46	47	47
11	49	49	49
12	51	51	51
13	53	52	52
14	54	54	54
15	55	56	56
16	58	58	58
17	61	61	61
18	63	64	64
19	67	67	66
20	69	69	68
21	70	70	70
22	72	71	72
23	72	73	73
24	73	74	74
25	74	75	76
26	77	77	78
27	81	79	80
28	83	82	83

4.7 Summary and Discussion

The purpose of linking is to establish the relationship between scores on two measures of closely related traits. The relationship can vary across linking methods and samples employed. In equipercentile linking, the relationship is determined based on the distributions of scores in a given sample. Although IRT-based linking can potentially offer sample-invariant results, they are based on estimates of item parameters, and hence subject to sampling errors. A potential issue with IRT-based linking methods is, however, the violation of model assumptions as a result of combining items from two measures (e.g., unidimensionality and local independence). As displayed in Figure 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 8 reports four statistics summarizing the current sample in terms of the differences between the PROMIS Depression T-scores and HADS Depression scores linked to the T-score metric through different methods. In addition to the seven linking methods previously discussed (see Figure 10), the method labeled "IRT pattern

Table 8: Observed vs. Linked T-scores

Methods	Correlation	Mean Difference	SD Difference	RMSD
IRT pattern scoring	0.712	0.336	8.435	8.438
IRT raw-scale	0.647	0.200	9.238	9.236
EQP raw-scale SM=0.0	0.655	0.143	9.292	9.289
EQP raw-scale SM=0.3	0.653	0.661	9.724	9.742
EQP raw-scale SM=1.0	0.650	0.788	9.871	9.898
EQP raw-raw-scale SM=0.0	0.656	0.078	9.245	9.242
EQP raw-raw-scale SM=0.3	0.655	0.050	9.224	9.220
EQP raw-raw-scale SM=1.0	0.654	0.002	9.247	9.243

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 (0.712), followed by EQP raw-raw-scale SM=0.0 (0.656). Similar results were found in terms of the standard deviation of differences and root mean squared difference (RMSD). IRT pattern scoring yielded smallest RMSD (8.438), followed by EQP raw-raw-scale SM=0.3 (9.22).

One approach to evaluating the robustness of a linking relationship is comparing the observed and linked scores in a new sample independent of the sample from which the linking relationship was obtained. Such a sample can be used to examine empirically the bias and standard error of different linking results. Because of the small sample size (N=1112), however, subsetting out a sample was not feasible. Instead, a resampling study was used where small subsets of cases (e.g., 25, 50, and 75) were drawn with replacement from the study sample (N=1112) over a large number of replications (i.e., 10,000).

Table 9 summarizes the mean and standard deviation of differences between the observed and linked T-scores by linking method and sample size. For each replication, the mean difference between the observed and equated PROMIS Depression T-scores was computed. Then the mean and the standard deviation of the means were computed over replications as bias and empirical standard error, respectively. As the sample size increased (from 25 to 75), the empirical standard error decreased steadily. At a sample size of 75, IRT pattern scoring produced the smallest standard error, 0.957. That is, the difference between the mean PROMIS Depression T-score and the mean equated HADS Depression T-score based on a similar sample of 75 cases is expected to be around $\pm 1.91 (i.e., 2 \times 0.957)$.

Examining a number of linking studies in the current project revealed that the two linking methods (IRT and equipercentile) in general produced highly comparable results. Some noticeable discrepancies were observed (albeit rarely) in some extreme score levels where data were sparse. Model-based approaches can provide more robust results than those relying solely on data when data is

Table 9: Comparison of Resampling Results

Methods	Mean_25	SD_25	Mean_50	SD_50	Mean_75	SD_75
IRT pattern scoring	0.332	1.644	0.320	1.169	0.338	0.957
IRT raw-scale	0.193	1.843	0.194	1.279	0.203	1.039
EQP raw-scale SM=0.0	0.155	1.830	0.145	1.281	0.146	1.033
EQP raw-scale SM=0.3	0.664	1.940	0.648	1.327	0.662	1.090
EQP raw-scale SM=1.0	0.779	1.961	0.810	1.363	0.792	1.115
EQP raw-raw-scale SM=0.0	0.073	1.832	0.068	1.281	0.092	1.024
EQP raw-raw-scale SM=0.3	0.053	1.827	0.048	1.273	0.046	1.030
EQP raw-raw-scale SM=1.0	-0.019	1.812	-0.015	1.270	0.000	1.031

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

5 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 Depression 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

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).

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.

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.

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

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

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.

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.

Zigmond, A. S., Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica 67* (6): 361-370.