

LINKING SCORES WITH PATIENT-REPORTED HEALTH OUTCOME INSTRUMENTS: A VALIDATION STUDY AND COMPARISON OF THREE LINKING METHODS

BENJAMIN D. SCHALET 

NORTHWESTERN UNIVERSITY, FEINBERG SCHOOL OF MEDICINE

SANGDON LIM 

THE UNIVERSITY OF TEXAS AT AUSTIN

DAVID CELLA

NORTHWESTERN UNIVERSITY, FEINBERG SCHOOL OF MEDICINE

SEUNG W. CHOI 

THE UNIVERSITY OF TEXAS AT AUSTIN

The psychometric process used to establish a relationship between the scores of two (or more) instruments is generically referred to as *linking*. When two instruments with the same content and statistical test specifications are linked, these instruments are said to be *equated*. Linking and equating procedures have long been used for practical benefit in educational testing. In recent years, health outcome researchers have increasingly applied linking techniques to patient-reported outcome (PRO) data. However, these applications have some noteworthy purposes and associated methodological questions. Purposes for linking health outcomes include the harmonization of data across studies or settings (enabling increased power in hypothesis testing), the aggregation of summed score data by means of score crosswalk tables, and score conversion in clinical settings where new instruments are introduced, but an interpretable connection to historical data is needed. When two PRO instruments are linked, assumptions for equating are typically not met and the extent to which those assumptions are violated becomes a decision point around how (and whether) to proceed with linking. We demonstrate multiple linking procedures—equipercentile, unidimensional IRT calibration, and calibrated projection—with the Patient-Reported Outcomes Measurement Information System Depression bank and the Patient Health Questionnaire-9. We validate this link across two samples and simulate different instrument correlation levels to provide guidance around which linking method is preferred. Finally, we discuss some remaining issues and directions for psychometric research in linking PRO instruments.

Key words: patient-reported outcomes, linking, scale alignment, depression, PROMIS, PHQ-9, calibrated projection.

1. Introduction

When two instruments measure similar constructs, it may be useful to establish a relationship between the scores obtained with the instruments. The psychometric process used to establish a relationship between the scores of two (or more) instruments is generically referred to as *linking* (Dorans 2007; Kolen and Brennan 2014). Linking can be further classified by the robustness of

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11336-021-09776-z>.

Correspondence should be made to Benjamin D. Schalet, Department of Medical Social Sciences, Northwestern University, Feinberg School of Medicine, 625 N Michigan Ave, 21st Floor, Chicago, IL60611, USA. Email: b-schalet@northwestern.edu

the resulting relationship, which is indicated in part by the correlation of the two instruments, along with several other linking properties (and supporting statistical indices). Chief among these properties are the extent to which instruments differ in the content or construct being measured, the measurement conditions (e.g., length, answer format, assessment administration procedures), and the population targeted by each instrument. If two linked instruments were developed according to the same content and statistical test specifications, as with alternate forms of a math test for example, the instruments are said to be *equated* (Kolen and Brennan 2014). Linking relationships that fail to meet this “same specifications” property are still commonly established and useful; an example of this would be the scale alignment between the ACT composite scores and the SAT Verbal-plus-Mathematics scores. This linking is described as a *concordance*, because the constructs are similar (but not the same) and the measurement conditions (such as test format and length) are different (Dorans et al. 1997).

Linking procedures have been applied in the development of educational tests for decades (Angoff 1971; Kolen and Brennan 2014; Lord 1980), and there have been increasing efforts to link patient-reported outcomes (PROs) measures in recent years. PROs may be defined as any report of the status of a patient’s health condition that comes directly from the patient (though caregiver reports may also be referred to as PROs). PROs assess patients’ perspectives on their symptoms, functional status, and quality of life (Basch 2014; Cella and Stone 2015). For example, PROs may assess clinically relevant psychological concepts, such as depression symptoms, and include self-reports of physical domains such as mobility and upper extremity function (Rose et al. 2014) or pain interference (Amtmann et al. 2010). They are frequently employed in clinical research, including clinical trials, to help evaluate treatment effectiveness from the patient’s perspective. PROs are also increasingly used to screen patients for possible follow-up, inform ongoing treatment, and monitor quality of treatment. Purposes for linking PRO instruments vary, but in the broadest sense it is to permit group-level comparisons and analyses on a single common metric.¹

1.1. Motivations for Linking with Patient-Reported Outcomes versus Educational Tests

The field in which PROs are used represents a somewhat different landscape than educational measurement and testing, where tests are centralized around specific programs or purposes, such as the SAT, ACT, the National Assessment of Educational Progress, or the Programme for International Student Assessment (e.g., von Davier et al. 2019). In contrast, new PRO instruments are developed and modified frequently, resulting in multiple assessment options for similar health domains. The multiplicity of these measures and the absence of a dominant standard instrument for any given health domain is a major impediment to scientific progress (Ahmed et al. 2012). One review of anxiety measures, for example, found 92 empirically based self-report instruments (McHugh et al. 2011). In addition, PRO instruments may differ widely in question format, length, number and type of response options, recall period, scoring method, clinically relevant benchmarks, and the amount of evidence for reliability and validity (De Vet et al. 2011). Furthermore, PRO measures differ greatly in the level of quality and rigor with which they were developed (Mokkink et al. 2010; Park et al. 2013; Uijen et al. 2012). The incommensurability of PRO scores across instruments can represent a problem for the downstream synthesis of individual studies, as it may not be clear whether differences in scientific findings are “real” or reflect methodological artifacts. Given this context, linking studies promise to provide bridges across separately developed instruments, enabling comparison of scores across samples and healthcare settings.

In educational testing the motivations for linking instruments can be quite different. Equating specifically can be important in a high-stakes testing environment, because it is key to maintaining test security and avoiding practice effects (Kolen and Brennan 2014). Equating assumptions must be met in order to create alternate forms or to develop new content to incorporate in an existing item pool. Test security and practice effects, however, are generally not of concern in the assessment of

PROs, because there is no “correct” answer to share with fellow test-takers, and patients are not manifestly motivated to achieve a particular score.² The goal of PRO-based linking is typically not to develop new alternate forms or expand an item pool, but rather to establish linkages between the large number of PRO instruments that measure sufficiently similar constructs.

There are various motivations for linking in the health outcomes field. One is the harmonization of PRO data across multiple longitudinal cohort studies, each with separate assessment protocols. This harmonization goal is important, because it would allow hypothesis testing to proceed on combined datasets. Individual cohort investigators will have historical data on similar constructs collected with different instruments and will understandably be constrained in their willingness and/or ability to change instruments in order to standardize across cohorts. There is a tension between maintaining standardization within a cohort of individuals followed longitudinally and the need to make measures comparable to those collected in other studies. Separate linking studies can provide solutions in this context (Kaat et al. 2017; Schalet et al. 2020).

A second prototypical need for linking is represented by the hospital director in charge of monitoring healthcare quality, which may include reporting on PRO *performance* measures (PRO-PM) for a particular medical care facility. This score may serve as one of the criteria in an incentive payment program (Basch et al. 2015). A hospital director may wish to switch from a costly proprietary PRO instrument to a highly similar, less expensive or more acceptable (e.g., briefer; easier to read) instrument. But if instruments are switched, it could mean that the connection to historical data is lost and comparisons with other healthcare facilities are no longer possible. Linking studies could build evidence and tools to switch to a preferred instrument, while still enabling the desired group-based comparisons, as demonstrated by Katzan et al. (2017) with depression measures.

Finally, linking also has potential utility in clinical contexts where PROs are used to monitor individual treatment progress. Although it is inadvisable to rely solely on a linked score to make an individual treatment decision (i.e., termination of treatment), this limitation does not apply to the group-level reference data that clinicians may want to access. For example, an orthopedic surgeon might compare an individual patient’s PRO pain interference score trajectory to a large group of patients with a similar fracture to help determine whether the patient is recovering normally (Baumhauer and Bozic 2016). But this goal cannot be realized if one instrument is used routinely in the clinic with individual patients (e.g., PROMIS Pain Interference; Amtmann et al. 2010), but summary scores from another (e.g., Brief Pain Inventory; Cleeland et al. 1994) are available as the patient reference population. Once these measures have been linked, the reference values could potentially be expressed in the metric of multiple instruments. The availability of these linked values (especially in graphical form) as reference points would avoid the need to double-up on assessments and also give clinicians more flexibility in the choice of PRO instrument to monitor treatment progress.

The above health outcome applications illustrate the potential wide-spread utility of linking and subsequent group-level analyses. This purpose distinguishes PRO applications from linking in high-stakes educational testing situations, where a critical decision might be made on the basis of an individual equated score or cutoff value (e.g., a college admission or scholarship award). When linking is undertaken to support high-stakes decisions, it is of greater importance that the linking meet stringent assumptions. For example, Dorans (2004) has argued that equating and concordance should only be conducted when the observed correlation between the scores of the instruments is 0.866 or higher. At this value, the reduction of uncertainty in the score metric is 50%, i.e., one minus the coefficient of alienation ($1 - \sqrt{1 - r^2}$). Linking studies with PRO instruments, however, are more likely to show raw summed score correlations between instruments in the 0.75 to 0.85 range (e.g., Choi et al. 2014; Cook et al. 2015; Kaat et al. 2020; Kaat et al. 2017; Lai et al. 2014; Schalet et al. 2014; ten Klooster et al. 2013; Tulskey et al. 2019).

1.2. PROMIS as a Standard Metric for Linking

A considerable number of linking studies in health outcomes have been conducted with Patient-Reported Outcomes Measurement Information System (PROMIS[®]) instruments. Because PROMIS instruments were developed using IRT models, and are routinely scored with IRT parameters, they lend themselves well toward forming a common metric for linking. The metric is additionally convenient, because the item parameters of most PROMIS item banks are centered on an estimate of the general population, such that a T-score of 50 represents the mean and 10 the standard deviation (Liu et al. 2010). PROMIS itself can be seen as an effort to address the vast heterogeneity of PROs by encouraging the field to adopt a common set of generic or universal health domains (Cella et al. 2007; Gershon et al. 2012; Reeve et al. 2007). PROMIS researchers developed and calibrated multiple item banks on a variety of physical, mental, and social health domains, such as pain (Amtmann et al. 2010), sleep (Buysse et al. 2010), depression and anxiety (Pilkonis et al. 2011; Pilkonis et al. 2014), physical function (Rose et al. 2014), and ability to participate in social roles (Hahn et al. 2014), among others.

PROMIS investigators standardized psychometric procedures for item bank development (Reeve et al. 2007), which were later expanded to illustrate multiple-group calibration, new approaches to DIF analysis, and CAT simulations (Hansen et al. 2014). Because PROMIS items for a single health domain are calibrated together, different combinations of PROMIS items within a bank can be scored on a common T-score metric, enabling group comparison of scores across test forms and studies (Cella et al. 2010; Segawa et al. 2020). PROMIS item banks are calibrated with the graded response model (GRM) and are scored, in standard practice, with *expected a posteriori* (EAP) scores transformed to T-score metric (Bock and Mislevy 1982; Samejima 1969). The development of PROMIS represented a shift in how health outcomes are measured, because the latent construct is conceptualized independently from the particular set of items. This feature of PROMIS in turn opens the door toward IRT-based linking between PROMIS and commonly used “legacy” instruments, which are defined as well-validated instruments developed before PROMIS that measure highly similar concepts (Cella et al. 2010).

1.3. PRO Linking Methodology

Linking studies with PRO and PROMIS instruments have included the assessment of a wide range of mental and physical health domains in both general and condition-specific populations. A summary of these studies is provided in **Online Resource 1**. While the methodologies of these studies vary, most follow unidimensional IRT linking methods, equipercenile linking, or both. We will outline the basic procedures followed in publications from the “PROsetta Stone” project (Choi et al. 2014; Cook et al. 2015; Lai et al. 2014; Schalet et al. 2014) and describe a multidimensional alternative approach (ten Klooster et al. 2013; Thissen et al. 2011).

The PROsetta Stone project was conceived to apply linking procedures to align scores on PROMIS measures with scores from other closely related legacy instruments. The project team applied a multi-method approach that would allow raw scores on a range of PRO instruments to be expressed as standardized PROMIS T-score metrics, totaling approximately 58 linkages across 18 health domains. These linkages are catalogued on www.prosettastone.org, with the most detailed published report by Choi et al. (2014). The methods were modeled on linking procedures in the educational measurement literature (Kolen and Brennan 2014; Lee and Lee 2018).

Design. The PROsetta Stone project relied on archival data (collected for a different purpose) as well as newly collected data.³ Linking with PROMIS measures followed what might be called a *single-group* linking design, because all participants received PROMIS items and all of the legacy scale items together. Because PROMIS items have established (operational) calibrations, the PROsetta linking design could also be classified as *common-item equating [i.e., linking] to a calibrated item pool* (Kolen and Brennan 2014). From this perspective, the PROMIS items

are the common items administered to two samples. The common items have calibrations from the original calibration sample and the new linking sample, while the legacy instrument items are only administered to the linking sample. A benefit of these designs is the ability to directly evaluate the differences between observed PROMIS scores and the estimated linked PROMIS scores. A competing educational linking design—random-groups—is used in the development of alternate test forms; to our knowledge, this design is not used for PRO-based linking, in part because the correlation between the two instruments is not known and is needed to determine whether assumptions for linking are met.

Evaluation of Linking Requirements. Two instruments are linked because they are known to measure the same or similar constructs. It would not be appropriate, for example, for an anxiety instrument to be linked to a fatigue instrument (Dorans 2007). This can be investigated by qualitatively examining the overlapping content of the items and quantitatively by calculating the raw-score correlation between the instruments. The PROsetta Stone team applied a minimum of Pearson $r = 0.70$, but correlations were typically in the 0.80–0.90 range (Cella et al. 2016). If IRT is to be applied, a series of analyses focused on unidimensionality and related assumptions should be conducted, which are not unique to linking of course, and are well described elsewhere (Reeve et al. 2007; Reise et al. 2013). Finally, it is important to investigate whether the linking relationship varies by subgroup. Sample size permitting, this can be explored by simply calculating whether standardized demographic or clinical group differences are invariant across instruments and computing a root mean square deviation (RMSD) for separate subgroup linking relationships (Dorans and Holland 2000).

Equipercntile Linking. In equipercntile linking, we calculate scores on both the linked measure and the PROMIS measure, along with each score's percentile rank within the sample. The scores of the two measures are then aligned by associating scores with equivalent percentile ranks on the two score distributions. An advantage of this method over simpler forms of linking (i.e., adjusting for the mean and variance) is that it accommodates differences between instruments that vary across the score continuum. A practical advantage of equipercntile linking over IRT is that it does not require item-level data; this approach may be useful for linking PROMIS CAT scores, for example, to the summed scores of a legacy instrument. The impact of random sampling error can be minimized by applying a smoothing algorithm (Brennan 2004; Reinsch 1967).

Fixed-Parameter Calibration. In fixed calibration, items from a legacy measure and the PROMIS items are calibrated in a single run with PROMIS item parameters fixed at their previously published values (Kim 2006). The item parameters of the legacy measure are freely estimated, subject to the metric defined by the PROMIS item parameters. Thus, this calibration yields item parameters for the legacy measure that are on the PROMIS metric.

Separate Calibration with Linking Constants. The second IRT-based method is separate calibration followed by the computation of transformation constants. This procedure uses the discrepancy between the established PROMIS item parameters and a newly calibrated estimation of PROMIS parameters to place the legacy items on the existing PROMIS metric. This avoids imposing the constraints inherent in fixed-parameter calibration (Kim 2006). It is referred as separate calibration because two separate sets of calibrations are needed for common items. PROMIS and legacy items are first calibrated concurrently (without fixing the PROMIS parameters). Second, the established PROMIS parameters are designated as the “anchor” to estimate multiplicative and additive constants needed to transform the newly calibrated PROMIS parameters to the metric of the established PROMIS parameters. Once we obtain the constants, we use them to linearly transform all the legacy instrument item parameters. Four procedures are commonly used to obtain the linking constants: mean/mean, mean/sigma, Haebara (1980) and Stocking-Lord (1983) The first two are based on the mean and standard deviation of the item parameters, while the latter two apply characteristic curve methods. In the Haebara approach, linking constants are found that minimize the sum of the squared differences between item characteristic curves, while the

Stocking-Lord approach finds the constants that minimize the squared differences between the test characteristic curves (TCCs).

Comparing IRT Linking Methods. The application of these IRT methods leads to multiple sets of parameters for the same items. To compare methods, one can examine the differences between the TCCs. If the differences between the expected summed score values are small (e.g., less than 1 score point), we may consider the methods interchangeable. In PROsetta Stone studies, the methods frequently converged and the fixed-parameter results were chosen, given the method's simplicity (Choi et al. 2014; Cook et al. 2015; Lai et al. 2014; Schalet et al. 2014). It is not uncommon for these two methods to produce similar results (Kang and Petersen 2012; Lee and Lee 2018). However, when the linking population is different from the calibration population, it may be advisable to choose the separate calibration method and to conduct an invariance analysis prior to proceeding.

From IRT Parameters to Raw Summed Score Equivalents. Once the preferred IRT parameters are selected, any pattern of responses on the legacy instrument can be scored with EAP pattern scoring (Bock and Mislevy 1982). However, this means that each summed legacy score is associated with a range of possible IRT scores, determined by the variety of response patterns. Given that not every user can implement IRT scoring and that item-level data may not be available, it is of practical benefit to assign a single most likely scale score (i.e., PROMIS T-score) to a given summed score from the legacy instrument. To do this, we apply the polytomous version of the Lord-Wingersky algorithm (Lord and Wingersky 1984; Thissen et al. 1995). Following the distributive law, this algorithm computes the probability of each successive response recursively on the basis of previously computed likelihoods, thereby reducing the computational burden. Details and examples of the algorithm are provided by Thissen et al. (1995) and Cai (2015, pp. 537–541).

Evaluation of Linking Method. The equipercentile linking method can be used as a check for possible violations of IRT assumptions (Kolen and Brennan 2014). If equipercentile and IRT methods converge, which was frequently the case in the PROsetta Stone project (Cella et al. 2016), it provides some reassurance. If not, equipercentile linking is preferred and could suggest that IRT parameter estimates are biased. We also evaluate the accuracy of each linking approach by comparing respondents' linked scores to their actual scores on the PROMIS metric. For each method, we compute correlations, as well as the mean difference, standard deviation of the differences in scores, and the root-mean square difference. Furthermore, plots of the difference scores across the score continuum can be used to further examine whether any of the methods introduce bias in different score ranges (Bland and Altman 1999; Carstensen 2010).

Alternative Methods. A potential weakness of the IRT methods described above is that they assume unidimensionality of the aggregate item set; that is, the latent correlation between the two instruments is assumed to be 1.0. Given that raw-score correlations typically range from .75 to .85, as noted above, linking results based on unidimensional IRT models may be inaccurate to differing degrees. To facilitate linking for instruments that may not measure exactly the same construct, Thissen et al. (2011) coined the term *calibrated projection* for a multidimensional method that models the correlation between two instruments. The term *projection* in the IRT literature refers to a linking method that predicts score distributions for one measure from the predictor scores on another measure, which are assumed to be error-free (Holland and Dorans 2006). In *calibrated projection*, however, the predictor scores are assumed to have measurement error; IRT is used to model the predictor score distribution.⁴

In this approach, a multidimensional item model is fitted to the item responses from the two instruments. The slope parameters for one instrument's items are estimated on an instrument-specific dimension, while fixed to 0.0 on the other dimension (and vice versa). For model identification purposes, the means and variances of each dimension are fixed to 0.0 and 1.0, respectively. The covariance between the two dimensions is freely estimated, along with the slopes and intercepts. When calibrated projection is applied to PROMIS and a legacy instrument, a multidimen-

sional GRM is used and PROMIS slope and intercept values need to be fixed to their previously calibrated values on the PROMIS-specific dimension. This ensures that the resulting parameters estimated on the other dimension (associated with the legacy instrument) can be used to generate scores on the PROMIS metric. Also, the mean and variance for the PROMIS dimension need to be freely estimated (rather than fixed to 0.0 and 1.0), so that they reflect the PROMIS metric, rather than the latent distribution of the linking sample. The mean and variance for the legacy instrument dimension are still fixed to 0.0 and 1.0. The model may also include additional path specifications or latent variables, to model (for example) local dependencies for very similar items across instruments.

The calibrated projection approach additionally requires a multidimensional modification of Lord and Wingersky's recursion algorithm to compute raw score crosswalk tables from a two-dimensional latent structure (Lord and Wingersky 1984; Thissen et al. 1995). PROMIS studies that linked pediatric and adult measures (Reeve et al. 2016; Tulsy et al. 2019) were based on calibrated projection and an extension for linear approximation (Thissen et al. 2015).

1.4. Outstanding Methodological Issues with PRO Linking

Benefit of Multidimensional Models. Several methodological challenges remain in the field of health outcome linking. One open question is whether multi-dimensional linking models are preferred over unidimensional IRT models and equipercentile methods. A multidimensional model may provide a better fit to the data, but the resulting linking relationship may not be different practically. ten Klooster et al. (2013) compared increasingly complex IRT models when empirically linking physical function measures and found little justification for the increased complexity, when judged by the level of agreement between observed and linked theta scores. Theoretically, the application of calibrated projection should improve the accuracy of linking relative to other methods, but this may also depend upon the level of correlation between the instruments.

Correlation of Measures to be Linked. The correlation between the two instruments to be linked is a key indicator for deciding whether to proceed with linking, and which method is appropriate. This value indicates whether IRT-based unidimensional linking will be successful and also provides (in addition to content review) an indication of whether the constructs being measures are the same or very similar. In the context of high-stakes educational testing, Dorans (2004) argued that an observed correlation of 0.866 is an appropriate lower-bound value for instruments that should be linked. In the context of group-level analysis, Choi et al. (2014) commented that this threshold might be lowered to a raw-score correlations of 0.75–0.80 for the linking of PRO instruments. It is unclear, however, what the consequences are of lowering this threshold for the aggregated error of the linked scores, and to what extent each of the methods we reviewed above contribute to this error.

Validation. The extent to which linking assumptions are met is also likely to indicate the replicability/validity of the linking relationship across different samples and populations. While a few studies incorporate cross-validation on separate samples (Fischer et al. 2012; Reeve et al. 2016; ten Klooster et al. 2013; Tulsy et al. 2019), most published reports on PRO instruments (such as the PROsetta Stone studies) are based on single samples. Liegl et al. (2016) found that re-estimation of depression item parameters is probably not necessary across multiple medical populations, such as primary care and orthopedic rehabilitation patients. The researchers suggest that variations in the parameter estimates are unlikely to be clinically meaningful for linking results at the scale and group level. While IRT calibrated linking relationships of PROs can indeed replicate across different samples (Askew et al. 2013; Cook et al. 2015), these findings should probably not be taken for granted (Fischer and Rose 2019; Kim et al. 2015).

The present study demonstrates scale alignment techniques with PROMIS instruments and addresses two issues of key importance: the validation (or replication) of linking relationships

across multiple samples and how the correlation between instruments relates to the accuracy of linking method. We examine the first issue by re-estimating linked PHQ-9 depression parameters obtained by Choi et al. (2014) in two separate (general population) samples. In this evaluation, we will also compute differences between observed and linked scores at the group level, rather than at the individual level. We examine the second issue by first applying calibrated projection to the empirical data and comparing the resulting linking relationship to equipercents and unidimensional IRT linking methods. We also calculate how changing only the correlation strength between the instruments would have affected the resulting linked relationships in calibrated projection. Finally, we simulate how well the linked scores (or crosswalk tables) derived from each method recover the true PROMIS score, as a function of correlation strength and linking method.

2. Methods

2.1. Patient-Reported Outcomes Instruments

PROMIS Depression. The PROMIS Depression bank v1.0 for adults consists of 28 items with a 7-day time frame and a 5-point response scale, with response options ranging from “Never” to “Always” (Cella et al. 2010; Pilkonis et al. 2011). Item content covers emotional, cognitive, and behavioral symptoms of depression, rather than somatic symptoms. As with most PROMIS measures, the items were calibrated under the graded response model (Samejima 1969) such that different subsets of items from the instrument bank can be reported on the same metric (Pilkonis et al. 2011), permitting CAT administration and short-form score comparison (Segawa et al. 2020). The T-score is the reporting metric of PROMIS Depression, transformed from the EAP estimate of θ , where $T\text{-score} = \theta \times 10 + 50$ (Bock and Mislevy 1982). Short-forms of 4-, 6-, and 8-items are also available (Cella et al. 2019) and may be scored from summed scores (Thissen et al. 1995).

Patient-Health Questionnaire-9. The PHQ-9 is a widely used nine-item instrument designed to assess depression in primary care (Kroenke et al. 2001). The PHQ-9 was developed to directly reflect the criteria for major depressive disorder in Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM-IV; American Psychiatric Association, 1994). Respondents rate their symptoms over the last 2 weeks, using a 4-point scale ranging from 0 (Not at all) to 3 (Nearly every day).

2.2. Samples Design and Participant Characteristics

We obtained the dataset for the original PHQ-9 link from www.healthmeasures.net, and we collected a second dataset for the purpose of linking validation. For each sample, adult participants were recruited by companies who maintain large internet panels. While these are convenience samples, they were roughly matched to US general population demographic targets of gender, age, race, and ethnicity. All participants completed all items. (No blocking was used.)

We refer to the sample used by Choi et al. (2014) for linking as the *Original* sample. This sample was provided by NIH Toolbox investigators, who collected data on 20 PROMIS items and the PHQ-9 for evaluation during the calibration phase of that project (Pilkonis et al. 2013). These data were collected by Greenfield Online (now Toluna; www.tolunagroup.com). Pilkonis et al. (2013) reported that participants passed a variety of validity checks (“red herring” questions) and respondents who took less than two seconds for any 10 items were removed. We refer to the sample we collected for the purpose of the present study as the *Validation* sample. To collect these data, we worked with a second internet panel company, www.op4g.com, who administered the full PROMIS bank (28 items) and the PHQ-9. Our Validation sample was also screened for “speeders”, defined as those who spent less than two seconds on average per question; 81 participants (4%)

TABLE 1.
Sample design details and demographic characteristics

Sample characteristic	Original sample	Validation sample
<i>N</i>	748	1810
Mean age in years (SD)	47.2 (15.2)	45.8 (17.4)
Gender (%)		
Male	43.9	45.6
Ethnicity (%)		
Hispanic	15.2	13.0
Race (%)		
White	80.1	74.2
Black/African American	9.1	12.3
Asian	2.8	5.4
Multiracial		2.9
Other	10.1	4.3
Education (%)		
Less than high school	4.8	5.7
High school diploma	27.1	26.7
Further educational attainment	68.0	67.6

The Original sample refers to the Toolbox sample used by Choi et al. (2014) for linking.

were removed for this reason. For the purposes of this study, we used 20 PROMIS items matched with the Original sample.

Table 1 provides the demographic characteristics of the two samples. Demographic proportions roughly conform to US Census distributions, with the exception of education, such that those without a high school diploma are underrepresented in these samples relative to the Census, a common finding (Hays et al. 2015; Liu et al. 2010).

2.3. Analysis Strategy

Our analysis focused on (1) validation of depression linking in two samples, and (2) a comparison of calibrated projection to unidimensional and equipercentile methods.

Validation of Linking. Following Choi et al. (2014), we first examined classical item statistics and the essential unidimensionality of the combined item sets with confirmatory factor analysis (CFA), applying these criteria: RMSEA < .08 (Browne and Cudeck 1992), TLI > .95, and CFI > .95 (Hu and Bentler 1998; Hu and Bentler 1999). To investigate whether subgroups are likely to require separate linking relationships, we compared whether standardized mean differences by sex and age were maintained across instruments. For the replicability analysis, we first conducted differential item functioning (DIF) analyses between the samples using ordinal logistic regression (Choi et al. 2011). Next, we conducted concurrent calibration of the items under graded response model (GRM: Samejima 1969), fixing the PROMIS items to established item parameter values (Kim 2006). The GRM defines the probability of endorsing the ordered response category $k = 0, 1, \dots, m_i$ of item i as

$$P_{ik}(\theta) \equiv P_{ik}^*(\theta) - P_{i(k+1)}^*(\theta),$$

where $(m_i + 1)$ is the number of response categories for item i , and the cumulative probability function $P_{ik}^*(\theta)$ is defined as

$$P_{ik}^*(\theta) \equiv \begin{cases} 1 & k = 0 \\ \{1 + \exp[-a_i(\theta - b_{ik})]\}^{-1} & 0 < k < m_i, \\ 0 & k = m_i \end{cases}$$

where a_i and b_{ik} are discrimination and category boundary parameters, respectively. We acknowledge that there are advantages associated with another unidimensional IRT approach, e.g., separate calibration followed by linear transformation with linking constants (Stocking and Lord 1983). As noted above, we have repeatedly found the results to be extremely similar to fixed-parameter calibration (Choi et al. 2014; Cook et al. 2015; Lai et al. 2014; Schalet et al. 2014). To simplify, we focus here only on fixed-parameter calibration results. Next, we computed crosswalk tables, mapping each summed score on the legacy measure to the PROMIS T-score (Lord and Wingersky 1984; Thissen et al. 1995).

The next step in the replicability study was to conduct equipercentile linking (Kolen and Brennan 2014; Lord 1982). The equipercentile method estimates a nonlinear linking relationship using percentile rank distributions of the two linking measures. Equipercentile linking proceeds by identifying scores on the target scale Y that have the same percentiles as scores on the input scale X , so that $e_Y(x) = F_Y^{-1}[F_X(x)]$, where F_X and F_Y are the cumulative distribution functions of X and Y , respectively, and F_Y^{-1} is the inverse of F_Y . While equipercentile linking typically only involves raw summed scores x and y , it may be used to map summed scores directly to scaled scores, such as θ values and PROMIS T-scores. To accomplish this, the scale Y summed scores must be converted to the θ metric using existing item parameters for scale Y , for example using the Lord and Wingersky (1984) recursion algorithm. Denoting the cumulative distribution function of the converted values as $F_{\theta(Y)}$, equipercentile linking identifies $e_{\theta(Y)}(x) = F_{\theta(Y)}^{-1}[F_X(x)]$. The equipercentile linking method can be used in conjunction with a presmoothing method such as the log-linear model (Hanson et al. 1994) or cubic-spline postsmoothing (Reinsch 1967) to reduce irregularities in score distributions or crosswalk tables that are due to sampling error. We applied log-linear presmoothing with six degrees, along with 1,000 bootstrap replications to estimate the standard error of equating (Albano 2016; Kolen and Brennan 2014).

Finally, for our validation study, we compared the IRT and equipercentile linking relationships obtained independently in each sample, and plotted these curves. Next, to gauge the impact, we computed the differences between observed PROMIS scores and those estimated by linked scores, using the original parameters obtained from the first sample (Choi et al. 2014). We considered T-score differences greater than a small effect (> 2 T-score points) potentially meaningful in this comparison. Studies establishing minimally important differences suggest that changes of 2 to 5 T-score points on the PROMIS metric are meaningful to patients (Jensen et al. 2017; Purvis et al. 2018; Revicki et al. 2008).

Comparison of Linking Methods with Empirical Data. In order to compare the crosswalk functions by linking method, we needed to apply calibrated projection (Thissen et al. 2011) to our empirical data. Calibrated projection proceeds by first fitting a multidimensional IRT (MIRT) model with a simple structure to a response dataset. For the PROMIS-PHQ-9 linking, a two-dimensional model is fitted to the item responses from the two measures with each dimension representing a measure. That is, the first set of items (i.e., PROMIS Depression items) loads on θ_1 and the slope parameters on θ_1 are fixed at their PROMIS anchor parameters, while fixed as 0.0 for θ_2 . The category boundary parameters for the PROMIS items are also fixed at their anchor parameters after a reparameterization (e.g., $c_k = -a \times b_k$). The fixing steps are needed because the linking sample is different from the original calibration sample from which the established

anchor parameter estimates were obtained. The second set of items (i.e., PHQ-9 items) loads on θ_2 and the slope parameters are freely estimated, while fixed at 0.0 for θ_1 . The mean and variance of θ_1 (i.e., PROMIS latent score) are freely estimated, while those for θ_2 (i.e., PHQ-9 latent score) are fixed at 0.0 and 1.0, respectively. The covariance between θ_1 and θ_2 is freely estimated.

With the two-dimensional MIRT model, each summed score on the second set of items (i.e., PHQ-9 items) can be mapped onto θ_1 , using a multidimensional extension of Lord–Wingersky recursion. Specifically, using the two-dimensional item parameters, the multidimensional Lord–Wingersky recursion algorithm can be used to obtain a likelihood value $L(x|\vec{\theta})$ at each two-dimensional grid point $\vec{\theta}$ for each summed score of PHQ-9, $x = 0, 1, \dots, 27$. Next, EAP estimates and covariances can be computed from the posterior values. The equations were adapted from a report on EAP estimation (Bryant et al. 2005). Given a p -dimensional vector $\vec{\theta}$, the p -dimensional EAP estimate given a summed score x is

$$E(\vec{\theta}|x) = \int_p \frac{\vec{\theta} L(x|\vec{\theta}) f(\vec{\theta}, \Sigma) d\vec{\theta}}{L(x|\vec{\theta}) f(\vec{\theta}, \Sigma) d\vec{\theta}},$$

which is approximated by

$$E(\vec{\theta}|x) = \sum \frac{\vec{\theta} L(x|\vec{\theta}) f(\vec{\theta}, \Sigma)}{L(x|\vec{\theta}) f(\vec{\theta}, \Sigma)},$$

where $L(x|\vec{\theta})$ is the previously computed likelihood of score level x given $\vec{\theta}$, and $f(\vec{\theta}, \Sigma)$ is a multivariate density value given the correlation matrix Σ estimated previously. The summation is taken over all $\vec{\theta}$ grid. Similarly, the $p \times p$ posterior variance–covariance matrix of $E(\vec{\theta}|x)$ is

$$C(\vec{\theta}|x) = \int_p \frac{C(\vec{\theta}) L(x|\vec{\theta}) f(\vec{\theta}, \Sigma) d\vec{\theta}}{L(x|\vec{\theta}) f(\vec{\theta}, \Sigma) d\vec{\theta}},$$

which is approximated by

$$C(\vec{\theta}|x) = \sum \frac{C(\vec{\theta}) L(x|\vec{\theta}) f(\vec{\theta}, \Sigma)}{L(x|\vec{\theta}) f(\vec{\theta}, \Sigma)},$$

where $C(\vec{\theta})$ is $[\vec{\theta} - E(\vec{\theta}|x)][\vec{\theta} - E(\vec{\theta}|x)]'$. Finally, the first-dimension component of $E(\vec{\theta}|x)$, which is θ_1 , is the projected PROMIS Depression score for the summed score x of PHQ-9. Likewise, the square root of the first value in the main diagonal of $C(\vec{\theta}|x)$ is the standard error $SE(\theta_1)$.

Comparison of Linking Methods with Simulation. We conducted a Monte Carlo simulation to assess how different levels of the latent correlation between instruments interact with different linking methods (equipercentile, unidimensional fixed calibration, and calibrated projection). We

further explore the relationship between the error variances of the original and linked (or projected) scores for each method. We first specified the instrument characteristics of PROMIS Depression and the PHQ-9 (e.g., number of items, response categories). The instrument correlation was varied in 0.95(−0.05)0.60 for a total of 8 levels. For unidimensional IRT and equipercentile methods, we also simulated a correlation level of $r = 1.0$. (For calibrated projection, $r = 0.95$ was the highest level used. Generating response data using $r = 1.0$ would make two-dimensional calibration difficult. Specifically, the estimated correlation would often exceed 1.0, which then makes running the Lord–Wingersky recursion impossible.)

The number of simulation trials was 20 in each level, which was sufficient to obtain a stable pattern. In each trial, 1000 two-dimensional θ vectors were sampled from a multivariate normal distribution with the specified correlation. The response dataset was generated from the established parameters for 20 PROMIS items and the estimated parameters for 9 PHQ-9 items (obtained via fixed calibration with the Validation dataset), as well as the two-dimensional θ values. For both equipercentile and unidimensional conditions, PHQ-9 item parameters were then estimated by performing a one-dimensional fixed-parameter calibration on the simulated response data, while PROMIS item parameters were fixed at their anchor values. For calibrated projection, item parameters in the PROMIS dimension were fixed to their anchor values, while freely estimating the item parameters in the PHQ-9 dimension. The mean and the variance of the latent PROMIS factor was freely estimated, while the mean and the variance of the latent PHQ-9 factor were fixed to be 0.0 and 1.0, respectively. For equipercentile score equating, log-linear presmoothing was performed on both the source scale and the destination scale (Hanson et al. 1994). For each trial, PHQ-9 raw summed scores were computed from the simulated response data. Using each crosswalk table produced from equipercentile linking, calibrated projections, and the unidimensional calibration, the PHQ-9 raw summed scores were mapped to their respective θ estimates for PROMIS Depression. The difference between the linked θ and the true PROMIS θ was used to compute root mean squared error (RMSE) values.

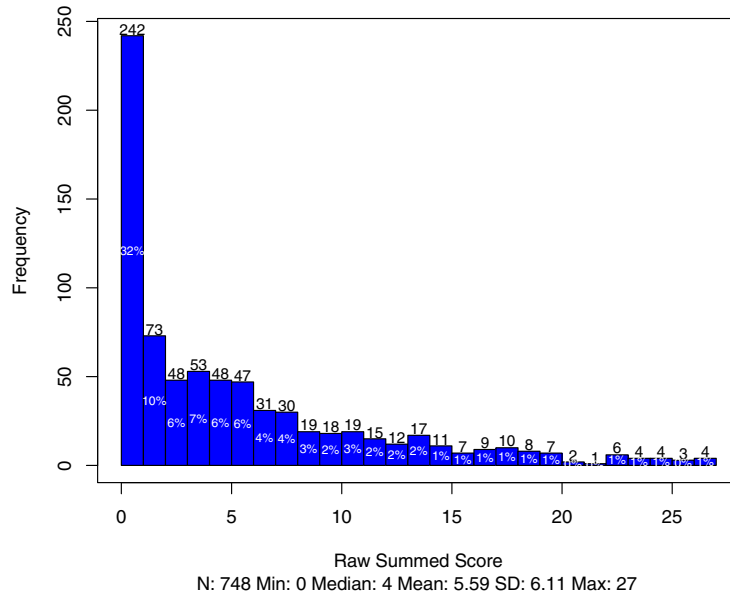
All analyses were conducted in R (v.4.0.2). We conducted CFA in lavaan (Rosseel 2012) and DIF analysis with lordif (Choi et al. 2011). Linking analyses were conducted with PROsetta (Choi et al. 2020), relying on equate (Albano 2016) for equipercentile linking and smoothing, and on mirt (Chalmers et al. 2012) for GRM parameter estimation, with NCYCLES modified to 1000 as the maximum number of EM cycles. **Online Resource 2** provides data and annotated R code to reproduce the linking and simulations results described below.

3. Results

3.1. Validation Analyses

Descriptive and Classical Statistics, Subgroup Differences. Figure 1a-b shows the distribution of raw summed scores across the PHQ-9 in each dataset. The PHQ-9 questionnaire in the Validation sample showed a bit higher level of depression ($M = 7.1$; $SD = 6.8$) than the Original sample ($M = 5.6$; $SD = 6.1$). For context, a PHQ-9 score of 10 or greater is indicative of moderate or worse depression symptoms (Kroenke et al. 2010). There is a considerable skew in the distribution, which is common for mental health symptom scores in non-clinical samples (Tomitaka et al. 2019). Raw summed score correlations between PROMIS and the PHQ-9 were relatively high in both samples: Original, $r = 0.84$; Validation, $r = 0.85$, with a disattenuated value of 0.89 for both. (We calculated the disattenuated correlation by dividing the raw-summed correlation by the square root of the product of their reliabilities, namely, Cronbach's alpha.) Classical item statistics seemed amenable to linking of PROMIS to PHQ-9 across the two samples. Briefly, the corrected item-total correlations ranged from 0.61 to 0.88 (Original sample) and 0.63 to 0.88 (Validation

(A) PHQ-9 Original Sample



(B) PHQ-9 Validation Sample

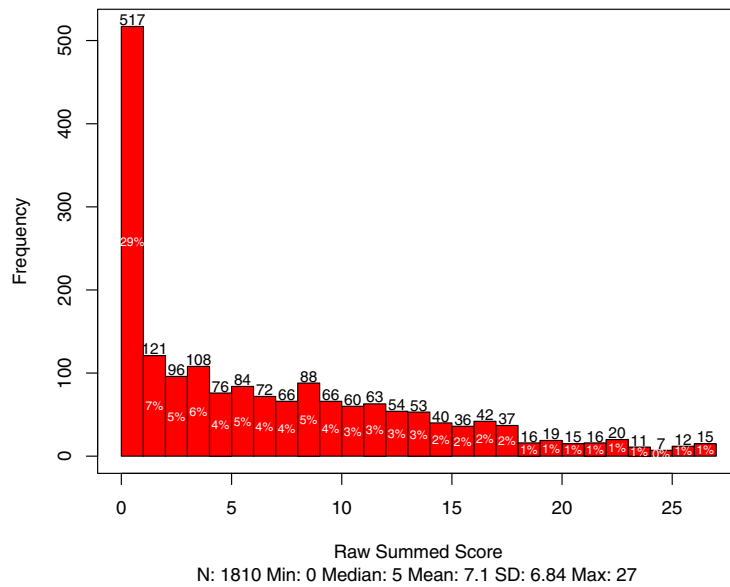


FIGURE 1.

Raw summed distributions for the PHQ-9 in Original and Validation samples **a** PHQ-9 Original sample, **b** PHQ-9 Validation sample

TABLE 2.
Depression score means, standard deviations, and standardized mean differences for sex and age by instrument

Sample	N*	Mean	SD	Standardized mean difference	
				Females – Males	Young (< 55) – Old (\geq 55)
<i>Original sample</i>					
PROMIS depression	748	49.98	9.73	0.259	0.268
PHQ-9	748	5.59	6.11	0.281	0.332
<i>Validation sample</i>					
PROMIS depression	1810	50.27	9.91	–0.005	0.626
PHQ-9	1810	7.10	6.84	0.009	0.653

PROMIS Depression scores are T-scores based on EAP pattern scoring. The Original sample refers to the Toolbox sample used by Choi et al. (2014) for linking.

sample). Table 2 shows the mean depression scores for each instrument by sample, as well as the standardized mean differences for sex and age. These differences were relatively consistent instruments, suggesting subgroup invariance.

Factor Analyses. For the combined item sets comprised of PROMIS items and the items of the legacy measures, values of single-factor CFA fit statistics were acceptable. For the Original linking set with PHQ-9, the combined item set (20 + 9 = 29 items) fit values were: CFI = 0.977, TLI = 0.975, and RMSEA = 0.087, 90% CI [0.084, 0.091]. We note that the latter value slightly exceeded the prespecified criteria of 0.08 (Browne and Cudeck 1992). For the Validation linking set with PHQ-9, the combined item set (28 + 9 = 37 items) fit values were: CFI = 0.977, TLI = 0.976, and RMSEA = 0.079, 90% CI [0.077, 0.080].

Differential Item Functioning Analysis. Differential item functioning analysis based on ordinal logistic regression was conducted between the Original and Validation samples for the PHQ-9 items (Choi et al. 2011). The likelihood-ratio χ^2 criterion ($\alpha = .01$) flagged 8 PHQ-9 items for either uniform or non-uniform DIF; however, all items had a pseudo R^2 change of less than .02. The PHQ-9 (suicidal thoughts) showed the largest amount of DIF, with uniform R^2 of .016, toward greater endorsement in the Validation sample.

Fixed Calibration to Obtain Linked Item Parameters. Collectively, the preceding procedures did not raise any red flags around proceeding with unidimensional linking. GRM parameters on the PHQ-9 were estimated fixing the PROMIS items to previously established item calibrations. Figure 2 shows the expected score curves for the linked parameters obtained from each PHQ-9 sample. Consistent with the pseudo R^2 change statistic, the expected score curves from fixed calibration demonstrate DIF in item #9. Figure 3 shows the two TCCs, as well as the information function and standard error curves (on the theta metric).

Crosswalk Tables. Figure 4a–c shows the relationship between the summed PHQ-9 scores and PROMIS T-score, as estimated by crosswalk tables derived from the unidimensional IRT, equipercentile, and calibrated projection methods, along with the marginal densities for each instrument's scores. The figures show how each sample produces a slightly different relationship, though this is most apparent at very severe levels of depression symptoms (T-scores > 75) for the equipercentile method, likely due to low sample sizes at this level in the Original sample. The unidimensional IRT and calibrated projection methods produced virtually the same linking relationship for each sample. (Please note that crosswalk tables by sample are available in **Online Resource 3**.)

Differences in Unidimensional IRT Linking across Samples. Choi et al. (2014) published a crosswalk table based on unidimensional IRT linking; we focus here on the differences between samples using that same method. Table 3 shows the differences between linked and observed

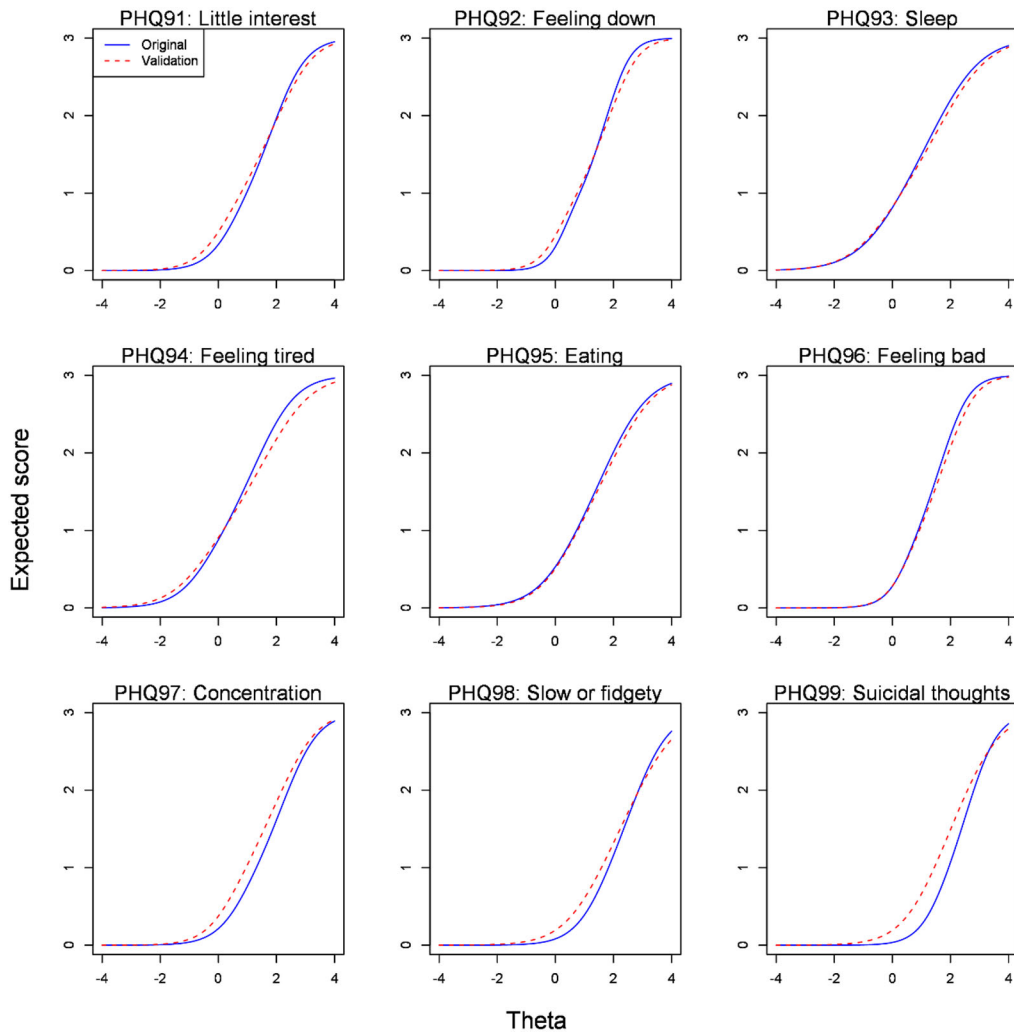


FIGURE 2.

Comparison of expected scores in Original and Validation linking analysis for PHQ-9. Item wording is truncated. Item PHQ99, “Thoughts that you would be better off dead, or thoughts of hurting yourself in some way?” showed the largest effect size for DIF (uniform + non-uniform)

PROMIS T-scores using the Validation sample. This was done by converting PHQ-9 scores in the Validation sample with PROMIS estimates from the crosswalk generated from the Original and the Validation sample. In addition, differences were explored by subgroup. For the PHQ-9 link, nearly all differences were very small (< 1 T-score point), with the highest being 1.35 T-score points for the subsample of participants with a high school degree or below.

3.2. Comparison of Linking Methods

Comparison of Linking Methods with Empirical Data. To compare the three linking methods (unidimensional IRT, equipercetile, and calibrated projection), we focus here on the Validation sample, which has the benefit of a larger sample size (1810 vs. 748). Figure 5 shows the crosswalk linking functions for each method. The three methods produced similar crosswalk functions,

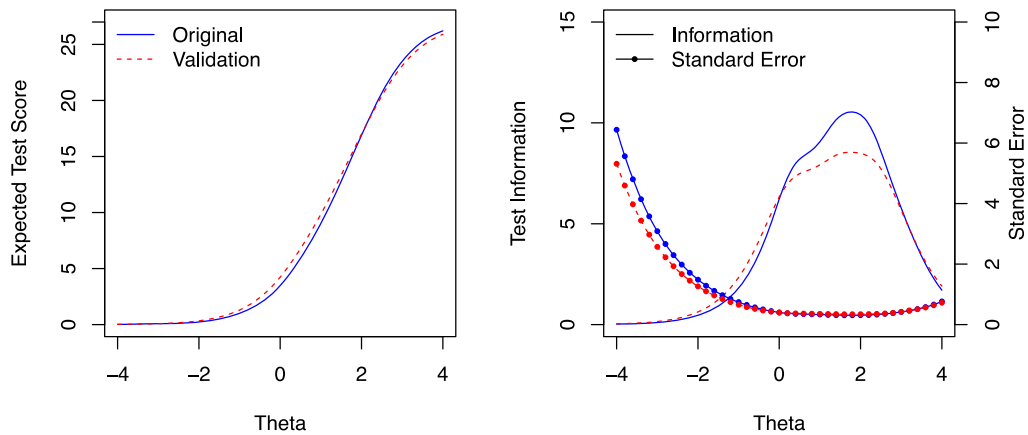


FIGURE 3.

Comparison of TCC and TIC in Original and Validation sample data for PHQ-9 using unidimensional IRT linking

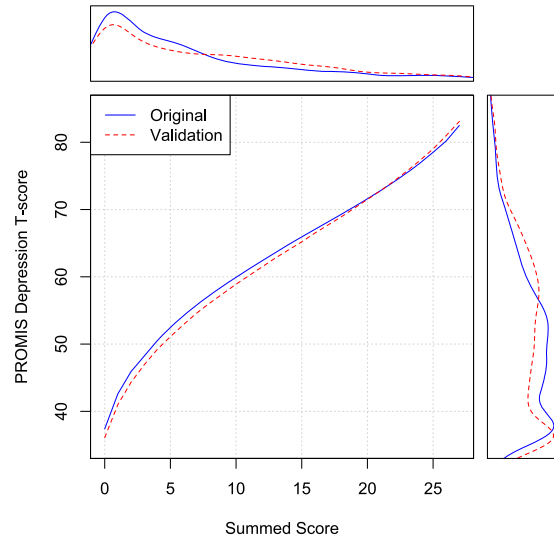
showing an average absolute deviation of 1.1 T-score points with one another (range 0.0–2.6). At high levels of depression (e.g., PHQ-9 of 20), calibrated projection linked to slightly lower PROMIS T-scores compared to the other two projection methods. For example, a PHQ-9 raw score of 20 links to a T-score of 69.6 with calibrated projection (SE = 6.4), but to a T-score of 71.5 with unidimensional IRT linking (SE = 3.5). As this example also shows, standard error estimates for calibrated projection model were (on average) 58% higher than for unidimensional IRT linking. Score crosswalk tables by method and standard error estimates are available in **Online Resource 3**.

The latent correlation between the measures was estimated to be 0.903 in the Validation sample. To investigate how different levels of the latent correlation would hypothetically affect the linked T-scores from calibrated projection, we re-ran calibrated projection on the response data, but used hypothetical values for the instrument correlation (0.7, 0.8, 0.9, and 0.99). To do so, we first needed to estimate the mean and variance for PROMIS only with unidimensional fixed calibration (fixing the item parameters to their previously established values). The estimated mean and variance were then used to additionally constrain the PROMIS factor mean and variance in the calibrated projection model, and to fix the PHQ-9-PROMIS covariance using our desired correlation value (0.7, 0.8, 0.9, and 0.99) multiplied by the square root of the PROMIS factor variance. (As before, the PROMIS item parameters were fixed to their previously calibrated values, and PHQ-9 factor mean and variance were fixed to 0.0 and 1.0.) To obtain the summed score crosswalk tables, we then ran the Lord–Wingersky recursion using each set of PHQ-9 item parameter estimates with the corresponding correlation values.

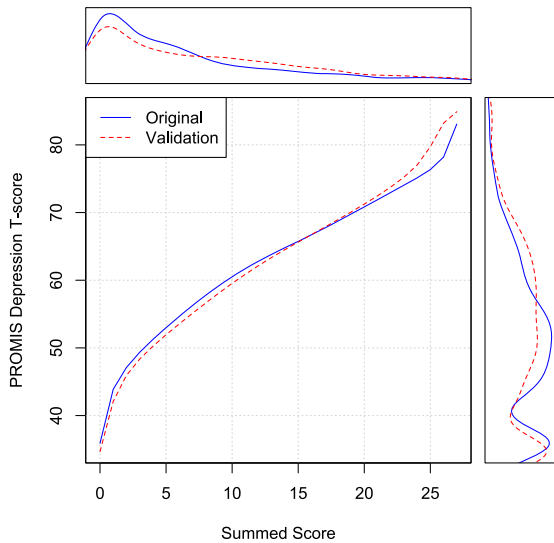
Figure 6 shows the result, namely calibrated projection at hypothetically different correlation levels, along with unidimensional and equipercentile relationships for comparison. The top right panel ($r = 0.90$) is equivalent to Fig. 5. Lower latent correlation levels, however, produce successively flatter curves, indicating that each raw score level in the PHQ-9 instrument was mapped to smaller variations of T-scores in the PROMIS instrument. At the extreme ($r = 0$), calibrated projection would map all raw score levels to a single T-score level because knowing the level of one instrument would not provide any information about the level of another instrument.

Comparison of Linking Methods with Simulated Data. In Fig. 7, we plot the RMSEs of the projected PROMIS Depression scores, θ_1 , by each of the three linking methods, across the different simulated latent correlation levels. As noted the level of $r = 1.0$ could not be simulated for calibrated projection, so this RMSE value was not available. Moving along the x -axis in

(A) Unidimensional IRT method



(B) Equipercentile method



(C) Calibrated projection method

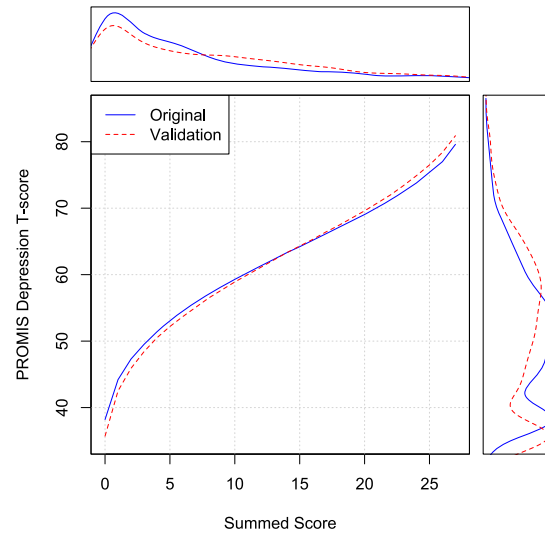


FIGURE 4.

Comparison of crosswalk tables based on Original versus Validation sample data for PHQ-9 summed score to linked PROMIS T-score **a** unidimensional IRT method, **b** equipercentile method, **c** calibrated projection method

TABLE 3.

Difference between linked and observed PROMIS T-scores, using crosswalk tables-based unidimensional IRT linking from Original versus Validation data

Validation sample characteristics and size	Observed PROMIS depression T-score			Difference between linked and observed T-scores in linking			
	<i>N</i>	Mean	<i>SD</i>	Original link		Validation link	
				Mean	<i>SD</i>	Mean	<i>SD</i>
Total sample	1810	52.33	12.77	0.68	6.90	-0.41	6.99
Female	984	52.32	12.62	0.80	6.80	-0.29	6.92
Male	826	52.34	12.95	0.53	7.00	-0.57	7.08
White	1390	52.24	12.76	0.74	6.83	-0.36	6.92
Black	248	53.42	12.57	0.51	7.89	-0.54	8.07
Hispanic	235	56.51	12.48	0.54	7.69	-0.32	7.83
High school and below	587	52.94	13.28	1.35	7.50	0.34	7.63
Some college and above	1223	52.04	12.51	0.36	6.56	-0.78	6.64
Younger (< 55)	1171	55.22	12.83	0.46	7.17	-0.49	7.31
Older (≥ 55)	639	47.04	10.82	1.08	6.36	-0.27	6.38

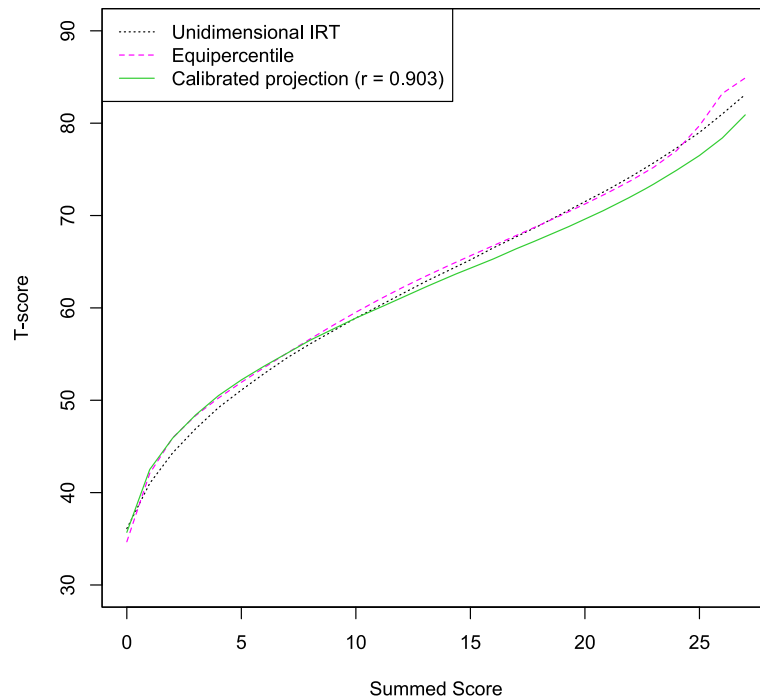


FIGURE 5.

Comparison of crosswalk tables for PHQ-9 summed scores to linked PROMIS T-scores by linking method (Validation sample)

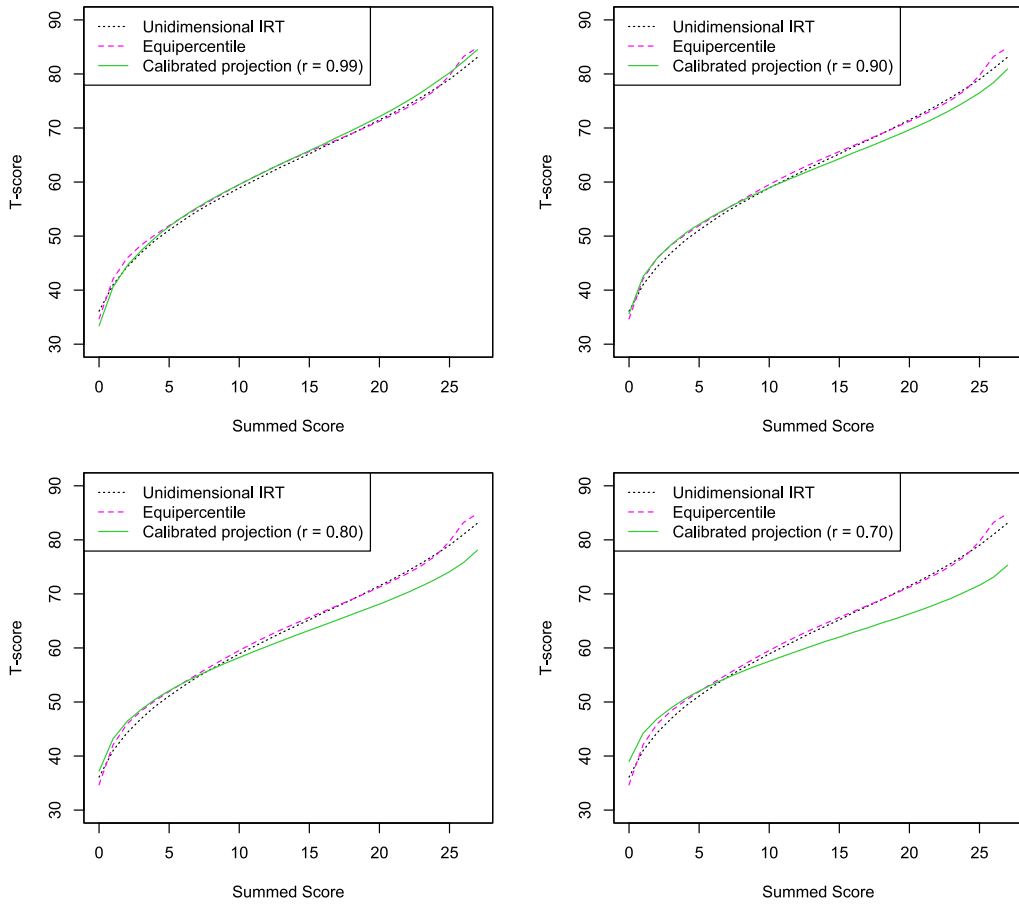


FIGURE 6.

Comparison of crosswalk tables for PHQ-9 summed scores to linked PROMIS T-scores with hypothetical values for the correlation between instruments in calibrated projection (Validation sample)

Fig. 7, we see that as the correlation decreases from an ideal linking situation the performance of the three methods naturally worsens but diverges progressively with calibrated projection outperforming the other two methods. The RMSEs for equipercentile and unidimensional IRT linking are indistinguishable until the correlation falls to 0.8 and diverges discernibly with the equipercentile linking performing slightly better than the unidimensional fixed calibration. Table 4 shows the corresponding values in the first three columns.

3.3. *Post-hoc Analysis of RMSE by correlation: Linear approximation for calibrated projection.*

Although the RMSEs for calibrated projection were obtained based on Monte Carlo simulations, it is possible to determine analytically how the RMSEs change as a function of the latent correlation, using linear approximation. Thissen et al. (2015) demonstrated that the projected scores and their error variances can be obtained by linear regression in lieu of two-dimensional numerical integration of the conditional posterior distribution (see Equations 1.6–1.14). The projected scores $\hat{\theta}_1$ can be obtained via a simple linear regression equation as

$$\hat{\theta}_1 = \beta_0 + \beta_1\theta_2, \tag{1}$$

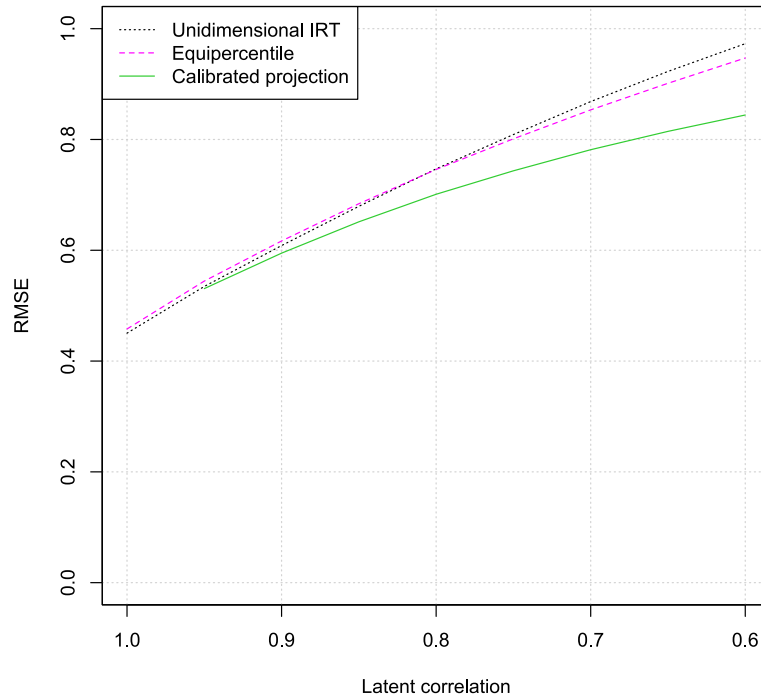


FIGURE 7.

RMSEs for unidimensional IRT, equipercetile, and calibrated projection linking methods across latent correlation levels (simulated response data)

in which the intercept β_0 and slope β_1 simplify to 0.0 and $\rho_{\theta_1\theta_2}$, respectively, when θ_1 and θ_2 are standardized each with mean 0.0 and SD 1.0. That is,

$$\hat{\theta}_1 = \rho_{\theta_1\theta_2}\theta_2. \quad (2)$$

Similarly, with the standardization, the conditional standard errors of the projected values can be simplified as

$$\widehat{SD}[\theta_1] = \sqrt{\rho_{\theta_1\theta_2}^2 SD^2[\theta_2] + (1 - \rho_{\theta_1\theta_2}^2)} = \sqrt{1 - \rho_{\theta_1\theta_2}^2 (1 - SD^2[\theta_2])}, \quad (3)$$

where $SD^2[\theta_2]$ is the error variance of the predicting values. The equation reveals that $\widehat{SD}[\theta_1]$ is necessarily larger than $SD^2[\theta_2]$ when $\rho_{\theta_1\theta_2}^2 < 1.0$, which is generally the case. In other words, the standard error of the projected scores, $\widehat{SD}[\theta_1]$, has a lower bound of $SD[\theta_2]$ and increases as $\rho_{\theta_1\theta_2}$ decreases from 1.0. In the current context, the standard error of the projected PROMIS Depression scores (θ_1) cannot be lower than the standard error of the predicting PHQ-9 scores (θ_2). The standard error of projected scores cannot go below the measurement error associated with the predicting scores.

When bias is zero or close to zero, it follows from Eqs. 1–3 that

$$\widehat{RMSE}[\theta_1] = \sqrt{1 - \rho_{\theta_1\theta_2}^2 (1 - \widehat{RMSE}^2[\theta_2])}. \quad (4)$$

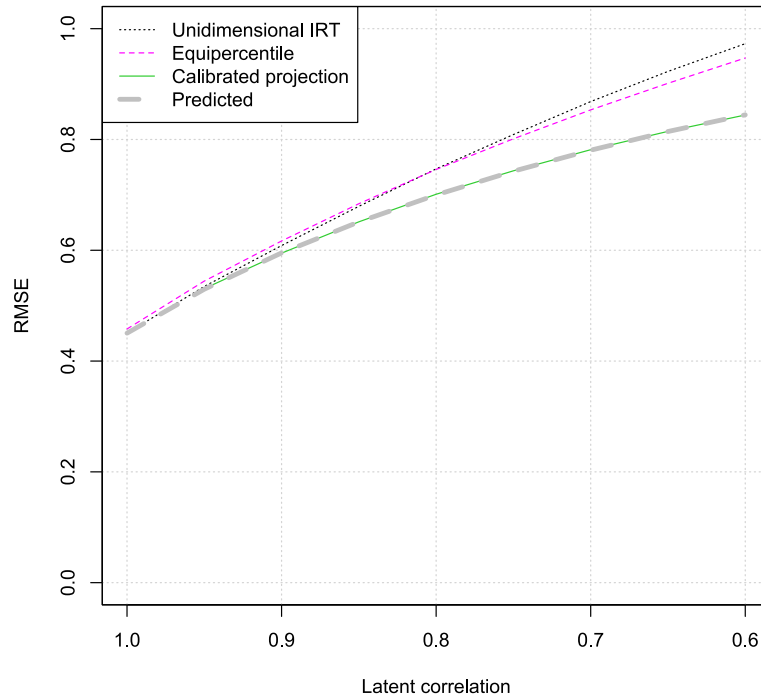


FIGURE 8.

RMSEs for unidimensional IRT, equipercntile, and calibrated projection linking methods across latent correlation levels (simulated response data), with prediction equation for calibrated projection

To approximate $\widehat{\text{RMSE}}^2[\theta_2]$, we can use 0.450, the RMSE value for unidimensional linking when $r = 1.0$. When $\rho_{\theta_1\theta_2}^2 = 0.9$, $\widehat{\text{RMSE}}[\theta_1] = \sqrt{1 - 0.9^2(1 - 0.450^2)} = 0.595$. When the correlation drops to 0.8, $\widehat{\text{RMSE}}[\theta_1] = \sqrt{1 - 0.8^2(1 - 0.450^2)} = 0.700$. Figure 8 shows that these analytically determined RMSEs (the thick dashed line labeled “Predicted”) almost perfectly match the RMSEs obtained from the calibrated projection Monte Carlo simulations. Table 4 shows the corresponding plotted values in the last column.

4. Conclusions and Future Research

4.1. Summary

Linking techniques offer much practical value to facilitate health outcomes research and clinical use of PRO instruments. Linking has the potential to effectively combine research findings across data repositories, maintain clinical interpretations across instruments, and facilitate harmonization, such that hypothesis testing can proceed on combined datasets. We demonstrated multiple linking procedures and added to this body of work by focusing on the validation of analyses for the linking between two of the most widely used depression symptom scales, PROMIS Depression and the PHQ-9 (Choi et al. 2014). Applying the original crosswalk—based on unidimensional IRT linking—to our data, sample and subgroup means differences were less than 1.5 T-score points. Calibrated projection produced an extremely similar crosswalk table when applied to both samples. Equipercntile linking, however, led to more discrepancies (between samples) at the higher end of the scale, likely due to the sparseness of data in the top end of the score

TABLE 4.

RMSE values for true PROMIS theta and linked PROMIS theta, estimated from PHQ-9, by different linking methods and latent instrument correlation

Latent correlation	Unidimensional IRT	Equipercetile	Calibrated projection	(Predicted)
1.00	0.450	0.458	–	0.450
0.95	0.535	0.544	0.531	0.530
0.90	0.608	0.617	0.595	0.595
0.85	0.679	0.684	0.651	0.651
0.80	0.747	0.746	0.701	0.700
0.75	0.809	0.801	0.743	0.743
0.70	0.868	0.853	0.782	0.781
0.65	0.923	0.901	0.815	0.814
0.60	0.973	0.947	0.844	0.844

RMSE = root-mean squared error. Values in the Predicted column are from Eq. 4, based on the linear approximation for calibrated projection (Thissen et al. 2015).

distribution. Equipercetile methods are known to be more sensitive to variations in the score distributions.

We also investigated how calibrated projection differs from unidimensional IRT and equipercetile linking at various levels of the latent trait correlation between instruments. Based on the empirical data of the PROMIS and PHQ-9 data, calibrated projection is quite similar to the other methods at a high latent correlation ($r = 0.90$), especially in the lower (non-clinical) range. But at a latent correlation of 0.80, calibrated projection produces substantially lower equivalent PROMIS T-scores (around 5 T-score points *lower* on depression) in the clinical range (T-score > 65), relative to the other two methods. Simulations additionally demonstrate that all three methods recover true PROMIS equivalently when latent correlations are 0.90 or greater. In fact, unidimensional linking should theoretically produce equivalent results to calibrated projection when $r = 1.0$ (following Eq. 2). However, calibrated projection does an increasingly better job of recovering true PROMIS theta as the correlation goes down. An interesting finding is that equipercetile linking performed slightly better than unidimensional IRT linking by the RMSE criterion. This is likely due to the greater violation of the unidimensionality assumption as the correlation between instruments decreases. Finally, we provide an equation that predicts the RMSE of the linked scores based on the latent correlation of two measures and the RMSE of the legacy instrument.

4.2. Validation of PRO Linking

An important remaining question is the extent to which PROMIS linking relationships further generalize to other populations, questionnaire forms, and data collection procedures. In the case of the PHQ-9 and PROMIS, we present evidence for validation of the original unidimensional IRT-based crosswalk, given a second sample selected from the general population. However, a few studies with patient data on both instruments have shown discrepancies, such that the PHQ-9 crosswalk over-estimates PROMIS at lower levels of depression (Katzan et al. 2017; Kim et al. 2015). Analyzing multiple sclerosis survey data, Kim et al. (2015) report similar levels of depression compared to our samples, but found that the PROMIS scores converted by the PHQ-9 crosswalk were on average 3.4 points higher (indicating more depression) than observed PROMIS scores. Following the same crosswalk table substitution, Katzan et al. (2017) reported an observed PROMIS T-score mean of 52.7 and a linked T-score of 55.3 for a large sample of neurology patients. It is notable that Katzan et al. (2017) report a Spearman correlation of 0.77,

which may reflect a slightly weaker association compared to the correlations we observed in the data we examined.

The above studies compared linked PROMIS T-scores to those from the 8-item PROMIS SF, whereas Choi et al. (2014) and our study were based on much longer PROMIS forms (20 and 28 items, respectively). The reliability of depression questionnaire measurement is low in the mild symptom range, regardless of the instrument used, but the longer PROMIS bank will mitigate this; for example, the lowest T-score with a model-based reliability of 0.90 is 39 for the full PROMIS bank (and 41 for the CAT), but 46 for the 8-item PROMIS SF (Segawa et al. 2020). Kim et al. (2015) note that the PHQ-9 linked PROMIS scores overestimate the level of depression on the PROMIS 8-item SF because a number of participants endorsed all of the lowest PROMIS response options (i.e., endorse no depression at all) but did not endorse the lowest PHQ-9 options. This raises the possibility that PROMIS 8-item SF has a higher floor compared to the PHQ-9. But an examination of the standard T-score tables of the PROMIS 8-item SF (www.healthmeasures.net) and the PHQ-9 crosswalk shows that the lowest scores are within 1 T-score point of each other.

Kim et al. (2015) also report a somewhat lower raw-score correlation ($r = 0.74$) between instruments compared to our samples ($r = 0.84$ and 0.85). All other factors being equal, our study illustrates that applying a crosswalk table from a unidimensional IRT method in a sample with a raw-score correlation of 0.74 (≈ 0.85 , when divided by square root of the product of the reliabilities) could indeed result in a mean difference between linked PROMIS and true PROMIS scores (see Fig. 6). Finally, as others have noted (Fischer and Rose 2019), the discrepancy Kim et al reported may simply be due sampling error (compounded by unreliability at the low end of the scale). Future linking with larger sample sizes can further clarify the role of sampling error. In addition, further analysis is needed to determine how different PROMIS forms may change the linking relationship, and whether extreme scores should be treated any differently.

4.3. Selection of Linking Method and Error

An informative aspect of our simulation work is that the RMSE between true theta and linked theta represents measurement error plus linking error. In this sense, the simulation answers the question of “how much” error is added when a legacy instrument (the PHQ-9 in our case) is used to estimate a linked PROMIS score. It demonstrates that linking error rises rather dramatically as the latent correlation goes down, regardless of the linking method used. For example, with unidimensional linking, an instrument correlation of 0.75 is associated with an increase in RMSE of about 80% compared the RMSE at $r = 1.0$, which can be conceptualized as PHQ-9 measurement error in the absence of linking error. The same comparison is associated with a 65% RMSE increase for the linear approximation of calibrated projection, suggesting some benefit to this method at $r = 0.75$. Equation 4 allows analysts to estimate and conceptualize the total error after linking with calibrated projection, given an instrument’s RMSE (and assuming little or no bias in θ estimation). This prediction may be useful tool to help analysts design linking studies.

Whether an increase in RMSE, relative to the RMSE at the “ideal” linking situation of ($r = 1.0$), represents a tolerable level of increased error is debatable and depends upon contextual factors that are driving the reasons for linking. For example, many applications of linking health outcomes (such as harmonization across datasets with different PRO instruments) are focused on group-level analyses of mean differences in large samples (Schalet et al. 2020); in this context an increase in variability due to linking may not matter, as long as there is no or little bias. However, if the results of a linking initiative will be used by a clinician to make an important treatment decision (e.g., surgery or physical therapy), it would be important to meet the minimum observed correlation threshold of .866 identified for high-stakes testing (Dorans 2004). We note that the overriding practical advice for the clinical use of any individual PRO score (linked or not) when

making critical decisions is to consider the standard error of the score (see, for example, Hays et al. 2005).

Our study provides guidance around the choice of which linking method to use, given an observed correlation between instruments. To compare this value fairly to the latent correlations presented in our simulation, we recommend disattenuating the correlation (using Cronbach's alpha for the reliability estimate).⁵ When the disattenuated correlation between instruments is 0.90 or higher, the results of each method are expected to converge and the differences among them will probably be inconsequential at the group level. In this circumstance, unidimensional IRT linking provides the benefit of creating a bi-directional link, as well as simplicity. Equipercile methods should be applied in tandem, to detect any cumulative effect of possible violations of IRT assumptions (Kolen and Brennan 2014). However, when disattenuated correlations are in the range of 0.75–.90, we recommend applying calibrated projection. Calibrated projection also has the benefit of modeling possible local dependencies or secondary factors that may have the potential to distort the linking relationship (by inflating discrimination parameters, for example). When calibrated project is not an option for instruments that correlate in this range, equipercile methods are preferred over unidimensional linking; this method also has the benefit of not requiring item-level data.

4.4. Limitations and Future Directions

We demonstrated linking procedures for two instruments that were highly correlated and reflect the relatively unidimensional construct of depression symptom severity (Pilkonis et al. 2011). We acknowledge that for instruments (and underlying constructs) with a greater degree of multidimensionality, a different linking strategy may be needed. PROMIS instruments were typically calibrated with a unidimensional model and scored using associated parameters, which limits the linking options. If the legacy instrument is multidimensional, choices need to be made around procedures, e.g., linking only to a particular subscale of the legacy instrument. Further research is needed to determine to what extent multidimensionality threatens the validity of the link, if this is ignored, and which method is best. For example, separate calibration followed by the application of linking constants may have an advantage in the presence of multidimensionality (Lee and Lee 2018).

Violations of the subgroup invariance assumption represent a threat to the validity of linking. In our analyses, we briefly checked that standardized mean differences for sex and age were similar. Measures that are highly correlated are likely to meet this assumption (Dorans and Holland 2000). However, the invariance assumption should be investigated for meaningful subgroups, ideally by repeating linking on subgroups and computing the differences between these results and those obtained for the larger sample. A more fine-grained approach is to examine DIF among subgroups (e.g., Kaat et al. 2017). The cumulative effect of DIF can then be examined, which might result in sample-specific parameters or linking relationships. But this diminishes the ease and utility of the results to score data in other settings and research; in addition, sample sizes would need to be large and researchers would have to have confidence in the rationale and replicability of the DIF findings. Troublesome DIF (or poorly fitting items) could also be deleted in order to estimate accurate parameters for linking. However, this decision needs to weigh the cost of excluding these items vs. the benefit of providing a summed score that represents the entire PRO measure.

The presence of multiple DIF groups and related covariates represents a possible challenge to the methodologies presented here. Moderated nonlinear factor analysis (MNLFA) has been offered as way to simultaneously analyze questionnaire items in multiple datasets, which are demonstrated in the context of alcohol and nicotine use questions (Curran and Hussong 2009; Gottfredson et al. 2019; Hussong et al. 2019; Rose et al. 2013). One benefit of this approach is that it integrates IRT and DIF analyses, making necessary adjustments for dichotomous and continuous subgroup

variables (such as age). In the context of PROMIS, however, this approach has the potential downside of producing factor scores adjusted by various DIF variables and covariates. These model-based factor scores would no longer be interpretable and comparable to PROMIS T-scores reported in the literature. The aim of MNLFA-based integrated data analysis is also somewhat differently conceptualized from linking as presented here; in this approach, parameter estimates and scoring are specific for a single harmonized dataset. The scoring results are not intended for general use in multiple samples or (potentially) multiple populations beyond those available to the analyst. Further research is needed on how linking approaches compare to MNLFA, and whether and under what conditions accounting for subgroups is necessary for linking PROMIS items and health outcome instruments.

Another limitation to our work pertains to the application of the GRM model to constructs that may not be normally distributed. Depression symptoms seem to follow what Reise and colleagues call a quasi-normal trait, where low levels of the trait are absent or irrelevant (Reise et al. 2018). While parameter estimates may have been biased, the similarity between IRT and equipercentile method in the empirical results suggests the problem of non-normality—in the context of linking—is limited. The application of a different model, however, such as the unipolar log-logistic model, could theoretically provide more information at low levels of the trait (Lucke 2014). For depression, it is arguably more important to make accurate decisions at the moderate range (e.g., referral to a psychiatrist), and yet aggregate quality measurement also depends upon the accuracy of determining if patients meet a mild symptom (remission) criteria (Katzan et al. 2017). Applying a zero-inflated model to our data would have probably led to lower (less biased) linked discrimination parameters and would have the benefit of maintaining the original GRM scoring metric (Wall et al. 2015). Further research is needed on the feasibility and possible improved linking accuracy of these approaches.

A final possible threat to the generalizability of linking is limited standardization in how PRO items are administered to respondents. While the PROMIS initiative has sought to minimize this variation, legacy measures and historical data may not adhere to these standards. For example, wording of item stems or response options may be changed without documenting these changes. While the preferred mode is to collect PROMIS data via unassisted self-report, sometimes PRO data are collected by a nurse or research assistant reading questions to patients in the clinic or over the phone. There are sometimes variations in how PRO items are presented to users, and these variations might be sources for study-based DIF and in turn introduce variation in parameter estimation or in the linking relationship. Even within a single mode (online administration), PRO administration variations exist: items may be presented one at a time or in a “grid” such that multiple questions appear on the page at once; questionnaire order may be randomized or not. Thus, a psychometrician tasked with linking, especially of historical data, needs to be attentive to these details and consider how these may affect results.

We presented an overview of linking methods with PROMIS measures, focusing on the PHQ-9 and PROMIS Depression. We replicated linking of Choi et al. (2014) using a large validation sample, and provided empirical and simulated comparisons of three linking methods, unidimensional IRT linking, equipercentile linking, and calibrated projection, as a function of instrument correlation. Our analyses should be repeated in clinical samples, and with multiple health domains, such as physical function, which may be more sensitive to patient characteristics (Coster et al. 2016). Psychometric work is needed to validate linking results to date by analyzing multiple (or larger) linking samples, across different populations, and to develop criteria for possible updating of linking parameters (Fischer and Rose 2019). Finally, the linking approaches presented here should be compared to different models and methods that accommodate a wider variety of DIF variables to evaluate the possible benefits and costs.

Endnotes

¹In this article, we use “linking” as an umbrella term, but in the framework of Holland and Dorans (2006), the PRO linkages enumerated in this paper are better described as a form of *scale alignment* (Holland, 2007). Scale alignment further has three designations: concordance, calibration, and vertical scaling, all of which, it could be argued, characterize the different PRO linkages we describe. *Concordance* requires that instruments purportedly measuring the same constructs have similar reliability, difficulty, and target a similar population. In many cases, however, the PROs being linked often have different reliabilities (lengths) and so this type of scale alignment is referred to as *calibration*, which also refers to the IRT procedure of estimating item parameters on a common scale, which may or may not occur during this type of scale alignment. Lastly, *vertical scaling* refers to linkages where measures have similar reliabilities but different difficulties and different intended populations. While this is less common, there is certainly impetus to do so in the health outcome field, such as when so-called generic and disease-specific PROs are linked (e.g., Victorson et al. 2019). In the domain of physical function, for example, one instrument may reliably measure disability while another measures a broader range of “activities of daily living” (e.g., Voshaar et al. 2019).

²An interesting exception to this is possible malingering or exaggeration of self-reported pain, in order to obtain pain medications. However, the conceptual basis and the reliable assessment of this phenomena are elusive (Tuck et al. 2019), a problem that cannot be solved with the development of alternate forms.

³The order of questionnaire administration should be counterbalanced, though for archival data, this may not be known. In educational testing this is especially important, because test-takers may not perform as well as fatigue sets in. We note that test-taking fatigue may be less of a concern for PRO questions because the questions are presumably cognitively less taxing than performance test questions. We suggest that counterbalancing is not as critical in PRO assessment.

⁴ten Klooster (2013) applied a method that is very similar to Thissen’s (2011) calibrated projection, applying a generalized partial credit model to link physical function instruments in a rheumatoid arthritis sample.

⁵Of course, a large disattenuated correlation value is necessary, but not sufficient to proceed with linking. Linking may not be possible or useful if instruments measure dissimilar constructs, exhibit low levels of reliability (< 0.80), or show poor indices of unidimensionality.

Acknowledgments

We wish to clarify that Seung W. Choi served as the senior author on this manuscript.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Ahmed, S., Berzon, R. A., Revicki, D. A., Lenderking, W. R., Moinpour, C. M., Basch, E., Reeve, B. B., Wu, A. W., & International Society for Quality of Life Research (2012). The use of patient-reported outcomes (PRO) within comparative effectiveness research: implications for clinical practice and health care policy. *Medical Care*, *50*(12), 1060–1070.
- Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*(8), 1–36.
- Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W.-H., Choi, S., Revicki, D., et al. (2010). Development of a PROMIS item bank to measure pain interference. *Pain*, *150*(1), 173–182.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.) *Educational measurement*. (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

- Askew, R. L., Kim, J., Chung, H., Cook, K. F., Johnson, K. L., & Amtmann, D. (2013). Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. *Quality of Life Research*, 22(10), 2769–2776.
- Basch, E. (2014). New frontiers in patient-reported outcomes: Adverse event reporting, comparative effectiveness, and quality assessment. *Annual Review of Medicine*, 65, 307–317.
- Basch, E., Spertus, J., Dudley, R. A., Wu, A., Chuahan, C., Cohen, P., et al. (2015). Methods for developing patient-reported outcome-based performance measures (PRO-PMs). *Value in Health*, 18(4), 493–504.
- Baumhauer, J. F., & Bozic, K. J. (2016). Value-based healthcare: Patient-reported outcomes in clinical decision making. *Clinical Orthopaedics and Related Research*, 474(6), 1375–1378.
- Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2), 135–160.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431–444.
- Brennan, R. (2004). Linking with Equivalent Group or Single Group Design (LEGS; Version 2.0)[Computer software]. Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment (CASMA).
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, 21(2), 230–258.
- Bryant, D. U., Smith, A. K., Alexander, S. G., Vaughn, K., & Canali, K. G. (2005). Expected a posteriori estimation of multiple latent traits (518612013-445)
- Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E., et al. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep*, 33(6), 781–792.
- Cai, L. (2015). Lord–Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, 80(2), 535–559.
- Carstensen, B. (2010). Comparing methods of measurement: Extending the LoA by regression. *Statistics in Medicine*, 29(3), 401–410.
- Cella, D., Choi, S. W., Condon, D. M., Schalet, B., Hays, R. D., Rothrock, N. E., et al. (2019). PROMIS® adult health profiles: Efficient short-form measures of seven health domains. *Value in Health*, 22(5), 537–544.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194.
- Cella, D., Schalet, B., Kallen, M., Lai, J.-S., Cook, K., Rutsohn, J., & Choi, S. (2016). PROSETTA stone analysis report: A rosetta stone for patient reported outcomes.
- Cella, D., & Stone, A. A. (2015). Health-related quality of life measurement in oncology: Advances and opportunities. *American Psychologist*, 70(2), 175.
- Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (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.
- Chalmers, R.P. mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29 (2012).
- Choi S, Lim S, Schalet B, Kaat A, & Cella, D. (2020). *PROsetta: Linking Patient-Reported Outcomes Measures*. R package version 0.2.0, <https://cran.r-project.org/package=PROsetta>
- Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, 39(8), 1.
- Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment*, 26(2), 513.
- Cleeland, C. S., Gonin, R., Hatfield, A. K., Edmonson, J. H., Blum, R. H., Stewart, J. A., et al. (1994). Pain and its treatment in outpatients with metastatic cancer. *New England Journal of Medicine*, 330(9), 592–596.
- Cook, K. F., Schalet, B. D., Kallen, M. A., Rutsohn, J. P., & Cella, D. (2015). Establishing a common metric for self-reported pain: Linking BPI pain interference and SF-36 bodily pain subscale scores to the PROMIS pain interference metric. *Quality of Life Research*, 24(10), 2305–2318.
- Coster, W. J., Ni, P., Slavin, M. D., Kisala, P. A., Nandakumar, R., Mulcahey, M. J., et al. (2016). Differential item functioning in the patient reported outcomes measurement information system pediatric short forms in a sample of children and adolescents with cerebral palsy. *Developmental Medicine and Child Neurology*, 58(11), 1132–1138.
- Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, 14(2), 81–100. <https://doi.org/10.1037/a0015914>.
- De Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press.
- Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4), 227–246.
- Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16(1), 85–94.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *ETS Research Report Series*, 2000(2), i–35.
- Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University*, 73(2), 24–32.
- Fischer, H. F., & Rose, M. (2019). Scoring depression on a common metric: A comparison of EAP estimation, plausible value imputation, and full Bayesian IRT modeling. *Multivariate Behavioral Research*, 54(1), 85–99.

- Fischer, H. F., Wahl, I., Fliege, H., Klapp, B. F., & Rose, M. (2012). Impact of cross-calibration methods on the interpretation of a treatment comparison study using 2 depression scales. *Medical Care*, *50*(4), 320–326.
- Gershon, R. C., Lai, J. S., Bode, R., Choi, S., Moy, C., Bleck, T., et al. (2012). Neuro-QOL: Quality of life item banks for adults with neurological disorders: item development and calibrations based upon clinical and general population testing. *Quality of Life Research*, *21*(3), 475–486.
- Gottfredson, N. C., Cole, V. T., Giordano, M. L., Bauer, D. J., Hussong, A. M., & Ennett, S. T. (2019). Simplifying the implementation of modern scale scoring methods with an automated R package: Automated moderated nonlinear factor analysis (aMNLFA). *Addictive Behaviors*, *94*, 65–73.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144–149.
- Hahn, E. A., DeWalt, D. A., Bode, R. K., Garcia, S. F., DeVellis, R. F., Correia, H., et al. (2014). New English and Spanish social health measures will facilitate evaluating health determinants. *Health Psychology*, *33*(5), 490.
- Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS® smoking item banks. *Nicotine and Tobacco Research*, *16*(Suppl 3), S175–S189.
- Hanson, B. A., Zeng, L., & Colton, D. A. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (Vol. 94). New York: American College Testing Program.
- Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K.-K. (2005). Evaluating the statistical significance of health-related quality-of-life change in individual patients. *Evaluation and the Health Professions*, *28*(2), 160–171.
- Hays, R. D., Liu, H., & Kapteyn, A. (2015). Use of Internet panels to conduct surveys. *Behavior Research Methods*, *47*(3), 685–690.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. *Educational Measurement*, *4*, 187–220.
- Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.
- Hussong, A. M., Gottfredson, N. C., Bauer, D. J., Curran, P. J., Haroon, M., Chandler, R., et al. (2019). Approaches for creating comparable measures of alcohol use symptoms: Harmonization with eight studies of criminal justice populations. *Drug and Alcohol Dependence*, *194*, 59–68. <https://doi.org/10.1016/j.drugalcdep.2018.10.003>.
- Jensen, R. E., Moinpour, C. M., Potosky, A. L., Lobo, T., Hahn, E. A., Hays, R. D. et al. (2017). Responsiveness of 8 Patient-Reported Outcomes Measurement Information System (PROMIS) measures in a large, community-based cancer study cohort. *Cancer*, *123*(2), 327–335.
- Kaat, A. J., Kallen, M. A., Nowinski, C. J., Sterling, S. A., Westbrook, S. R., & Peters, J. T. (2020). PROMIS® pediatric depressive symptoms as a harmonized score metric. *Journal of Pediatric Psychology*, *45*(3), 271–280.
- Kaat, A. J., Newcomb, M. E., Ryan, D. T., & Mustanski, B. (2017). Expanding a common metric for depression reporting: linking two scales to PROMIS® depression. *Quality of Life Research*, *26*(5), 1119–1128.
- Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review*, *13*(2), 311–321.
- Katzan, I. L., Fan, Y., Griffith, S. D., Crane, P. K., Thompson, N. R., & Cella, D. (2017). Scale linking to enable patient-reported outcome performance measures assessed with different patient-reported outcome measures. *Value in Health*, *20*(8), 1143–1149.
- Kim, J., Chung, H., Askew, R. L., Park, R., Jones, S. M., Cook, K. F., & Amtmann, D. (2015). Translating CESD-20 and PHQ-9 scores to PROMIS depression. Assessment, 1073191115607042.
- Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Berlin: Springer.
- Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.
- Kroenke, K., Spitzer, R. L., Williams, J. B., & Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. *General Hospital Psychiatry*, *32*(4), 345–359.
- Lai, J.-S., Cella, D., Yanez, B., & Stone, A. (2014). Linking fatigue measures on a common reporting metric. *Journal of Pain and Symptom Management*, *48*(4), 639–648.
- Lee, W. C., & Lee, G. (2018). *IRT linking and equating* (pp. 639–673). The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development.
- Liegl, G., Wahl, I., Berghöfer, A., Nolte, S., Pieh, C., Rose, M., et al. (2016). Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation. *Journal of Clinical Epidemiology*, *71*, 25–34.
- Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., et al. (2010). Representativeness of the patient-reported outcomes measurement information system internet panel. *Journal of Clinical Epidemiology*, *63*(11), 1169–1178.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. London: Routledge.
- Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, *7*(3), 165–174.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement*, *8*(4), 453–461.
- Lucke JF (2015). Unipolar item response models. In Reise SP & Revicki DA (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (pp. 272–284). New York, NY: Routledge/Taylor & Francis Group.

- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549.
- McHugh, R. K., Rasmussen, J. L., & Otto, M. W. (2011). Comprehension of self-report evidence-based measures of anxiety. *Depression and Anxiety, 28*(7), 607–614.
- Park, T., Reilly-Spong, M., & Gross, C. R. (2013). Mindfulness: A systematic review of instruments to measure an emergent patient-reported outcome (PRO). *Quality of Life Research, 22*(10), 2639–2659.
- Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS®): Depression, anxiety, and anger. *Assessment, 18*(3), 263–283.
- Pilkonis, P. A., Choi, S. W., Salsman, J. M., Butt, Z., Moore, T. L., Lawrence, S. M., et al. (2013). Assessment of self-reported negative affect in the NIH Toolbox. *Psychiatry Research, 206*(1), 88–97.
- Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., & Lawrence, S. M. (2014). Validation of the depression item bank from the patient-reported outcomes measurement information system (PROMIS®) in a three-month observational study. *Journal of Psychiatric Research, 56*, 112–119.
- Purvis, T. E., Neuman, B. J., Riley, L. H. I. I., & Skolasky, R. L. (2018). Discriminant ability, concurrent validity, and responsiveness of PROMIS health domains among patients with lumbar degenerative disease undergoing decompression with or without arthrodesis. *Spine, 43*(21), 1512–1520.
- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care, 45*(5), S22–S31.
- Reeve, B. B., Thissen, D., DeWalt, D. A., Huang, I.-C., Liu, Y., Magnus, B., et al. (2016). Linkage between the PROMIS®-pediatric and adult emotional distress measures. *Quality of Life Research, 25*(4), 823–833.
- Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik, 10*(3), 177–183.
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2013). Applying unidimensional item response theory models to psychological data. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbooks in psychology®. APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (p. 101–119). American Psychological Association.
- Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative approaches to addressing non-normal distributions in the application of IRT models to personality measures. *Journal of Personality Assessment, 100*(4), 363–374.
- Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*(2), 102–109.
- Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. (2013). An integrated data analysis approach to investigating measurement equivalence of DSM nicotine dependence symptoms. *Drug and Alcohol Dependence, 129*(1–2), 25–32.
- Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E. (2014). The PROMIS physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology, 67*(5), 516–526.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometrika Monograph Supplement No. 17) Richmond, VA Psychometrics Society.
- Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of Anxiety Disorders, 28*(1), 88–96.
- Schalet, B. D., Janulis, P., Kipke, M. D., Mustanski, B., Shoptaw, S., Moore, R., et al. (2020). Psychometric Data Linking Across HIV and Substance Use Cohorts. *AIDS and Behavior, 24*, 3215–3224.
- Segawa, E., Schalet, B., & Cella, D. (2020). A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administered using PROMIS profile. *Quality of Life Research, 29*(1), 213–221.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.
- ten Klooster, P. M., Voshaar, M. A. O., Gandek, B., Rose, M., Bjorner, J. B., Taal, E., et al. (2013). Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment Questionnaire disability index in rheumatoid arthritis. *Health and Quality of Life Outcomes, 11*(1), 1.
- Thissen D., Liu Y., Magnus B., Quinn H. (2015) Extending the Use of Multidimensional IRT Calibration as Projection: Many-to-One Linking and Linear Computation of Projected Scores. In van der Ark L., Bolt D., Wang WC., Douglas J., Chow SM. (Eds.), *Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics, vol 140* (pp 1–16). Springer, Cham.
- Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39–49.
- Thissen, D., Varni, J. W., Stucky, B. D., Liu, Y., Irwin, D. E., & DeWalt, D. A. (2011). Using the PedsQL™3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). *Quality of Life Research, 20*(9), 1497–1505.
- Tomitaka, S., Kawasaki, Y., Ide, K., Akutagawa, M., Ono, Y., & Furukawa, T. A. (2019). Distribution of psychological distress is stable in recent decades and follows an exponential pattern in the US population. *Scientific Reports, 9*(1), 1–10.

- Tuck, N. L., Johnson, M. H., & Bean, D. J. (2019). *You'd better believe it: The conceptual and practical challenges of assessing malingering in patients with chronic pain. The Journal of Pain, 20*(2), 133–145.
- Tulsky, D. S., Kisala, P. A., Boulton, A. J., Jette, A. M., Thissen, D., Ni, P., et al. (2019). Determining a transitional scoring link between PROMIS® pediatric and adult physical health measures. *Quality of Life Research, 28*(5), 1217–1229.
- Uijen, A. A., Heinst, C. W., Schellevis, F. G., van den Bosch, W. J., van de Laar, F. A., Terwee, C. B., et al. (2012). Measurement properties of questionnaires measuring continuity of care: A systematic review. *PloS One, 7*(7), e42256.
- Victorson, D., Schalet, B. D., Kundu, S., Helfand, B. T., Novakovic, K., Penedo, F., et al. (2019). Establishing a common metric for self-reported anxiety in patients with prostate cancer: Linking the Memorial Anxiety Scale for Prostate Cancer with PROMIS Anxiety. *Cancer, 125*(18), 3249–3258.
- von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., et al. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy and Practice, 26*(4), 466–488.
- Voshaar, M. O., Vonkeman, H., Courvoisier, D., Finckh, A., Gossec, L., Leung, Y., et al. (2019). Towards standardized patient reported physical function outcome reporting: Linking ten commonly used questionnaires to a common metric. *Quality of Life Research, 28*(1), 187–197.
- Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement, 39*(8), 583–597.

Manuscript Received: 14 JUN 2020

Final Version Received: 3 MAR 2021

Accepted: 19 MAY 2021

Published Online Date: 26 JUN 2021