Published in final edited form as:

Qual Life Res. 2017 May ; 26(5): 1119–1128. doi:10.1007/s11136-016-1450-z.

Expanding a common metric for depression reporting: Linking two scales to PROMIS® Depression

Aaron J. Kaat, Michael E. Newcomb, Daniel T. Ryan, and Brian Mustanski

Northwestern University

Abstract

PURPOSE—Depression is a significant mental health concern. There are numerous depression questionnaires, several of which can be scored onto the Patient-Reported Outcomes Measurement Information System (PROMIS®) Depression metric. This study expands the unified metric by linking depression subscales from the Adult Self Report (ASR) and the Brief Symptom Inventory (BSI) to it.

METHODS—An online sample of 2,009 men who have sex with men (MSM) was recruited. Item factor analysis was used to evaluate the dimensionality of the aggregated measures and confirm the statistical assumptions for linking. Then, linking was conducted using equipercentile and item response theory (IRT) methods. Equipercentile linking considered varying degrees of post-smoothing. IRT-based linking used fixed anchor calibration and separate calibration with Stocking-Lord linking constants.

RESULTS—All three scales were broadly unidimensional. This MSM sample had slightly higher average depression scores than the general population (mean = 54.4, SD = 9.6). Both linking methods provided robust, largely comparable results. Subgroup invariance held for age, race, and HIV status. Given the broad comparability across methods, the crosswalk between raw sum scores and the unified T-score metric used fixed anchor IRT-based methods.

CONCLUSIONS—PROMIS provides a unified, interpretable metric for depression reporting. The results of this study allow the depression subscales from the ASR and BSI to be rescored onto the unified metric with reasonable caution. This will benefit epidemiological projects aggregating data across various measures or time points.

Keywords

PROMIS; Depression; Item Response Theory; Linking

Correspondence concerning this article should be addressed to Aaron J. Kaat, Department of Medical Social Sciences, Northwestern University, 625 N. Michigan Avenue Suite 2700, Chicago IL, 60611. aaron.kaat@northwestern.edu; Phone 312-503-3477; Fax 312-503-4800.

Aaron J. Kaat, Department of Medical Social Sciences, Northwestern University; Michael E. Newcomb, Department of Medical Social Sciences, Northwestern University; Daniel T. Ryan, Department of Medical Social Sciences, Northwestern University; Brian Mustanski, Department of Medical Social Sciences, Northwestern University.

Conflict of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent: Informed consent was obtained from all individual participants included in the study.

Depression is one of the most common mental health concerns among adults. The 12-month prevalence for Major Depressive Disorder (MDD) among adults is approximately 7%, with a lifetime prevalence rate of approximately 16% [1]. However, debate persists as how to best quantify depression. Clinicians and researchers are faced with a myriad of patient-reported outcome (PROs) measures, each with unique measurement properties. The lack of standardization, for depression and other physical and mental health-related outcomes, was an impetus for the National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS®) [2]. PROMIS measures were developed using thorough qualitative item review and modern statistical methods—namely item response theory (IRT)—which allows for an interchangeability of items, computerized adaptive testing (CAT), reduced test burden, and many other measurement benefits. [2–4]

PROMIS Depression [5] is a recommended Level-2 Cross-Cutting Symptom emerging measure in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) [6]. However, multiple depression measures will persist across existing data, ongoing studies, and due to researcher preference. As such, it benefits the research and clinical community to be able to assess participants on different scales but have comparable scores. The metric is more important than the measures. Interest in a unified metric in depression has rapidly increased, especially as healthcare quality initiatives begin to address response and remission from depression. A unified metric “links” scores from different questionnaires onto the same scale. The scale chosen for PROMIS was the T-score (mean=50, SD=10) standardized to the 2000 USA general census in terms of gender, age, race, and education [7]. This allows comparisons to the general adult population, and thus is a prime candidate for a unified depression metric.

Gibbons and colleagues [8] were one of the first groups to capitalize on the PROMIS metric for depression. They linked the Patient Health Questionnaire (PHQ-9) [9] to the PROMIS metric in a sample of 2,178 HIV clinic participants (80% male) using fixed-anchor calibration. In a separate study [10], the PHQ-9, Beck Depression Inventory-II (BDI-II) [11], and the Center for Epidemiological Studies Depression Scale (CES-D) [12] were linked to the unified metric in a general population sample

PROMIS Depression has a separate pediatric bank (ages 5–17 for parent-report and 8–17 for child self-report), which is not centered on the same sample as the adult bank. Using a relatively new linking method, Reeve and colleagues [13] linked several self-report pediatric PROMIS measures, including depression, to their adult counterparts.

Two other linking studies are noteworthy, both of which linked PROMIS depression to other measures on a common metric, but it was not the PROMIS metric. The first study was conducted with a mixed adolescent and young adult sample [14]. The researchers used the adult bank, regardless of age, and co-calibrated PROMIS, the original BDI [15], and the CES-D [12]. The second study, and perhaps the largest linking study to date, was conducted by Wahl and colleagues [16]. Using a German clinical sample, they linked 14 scales from 11 measures to the same metric. Although neither of these studies used the PROMIS metric, their efforts support the goal of common score reporting.

The current study aimed to link the depression syndrome scale from the Adult Self Report (ASR) [17] and the depression subscale from the Brief Symptom Inventory (BSI) [18] to the PROMIS Depression item bank. Unlike most of the previous linking studies, which have been conducted on existing data collected for other purposes, the current study prospectively collected data for the purpose of linking the ASR and BSI to PROMIS. The results of this linking study will benefit the RADAR study, which is a longitudinal cohort of young men who have sex with men (MSM) that aims to examine multilevel influences on HIV and substance use. RADAR was built in part by merging together multiple prior longitudinal studies of young MSM (Project Q2 and Crew 450), each of which utilized different measures of depression [19; 20]. In order to maximize all previously collected data on depression prior to enrollment in the RADAR cohort, it was necessary to link the depression subscale from the BSI (Project Q2, five prior years of longitudinal data) and the ASR (Crew 450, two prior years of longitudinal data) with the PROMIS Depression measure that is being utilized in the RADAR study. Doing so will allow us to examine development trajectories of depression over time with up to 10 years of longitudinal data, as well as analyze how depression co-occurs with and/or predicts risk for HIV infection and substance use.

Method

Participants

We recruited participants via banner and pop-up advertisements placed on a geospatial smartphone application for men seeking men. The campaign served the dual purpose of recruiting participants for a randomized controlled trial (RCT; reported elsewhere) and to collect survey data from MSM that were ineligible for the RCT or all MSM once RCT recruitment targets were met.

Advertisements ran from November 2014 through February 2015 and described a university survey that provided an opportunity seeking input to better understand and serve the health needs of the lesbian, gay, bisexual, transgender, and queer (LGBTQ) community. Advertisements were shown throughout the USA, with pop-up ads shown up to five times—shown the first time a user logged onto the application within the scheduled 24-h advertising period. Banner advertisements ran continuously during the period. No incentives for participation were provided for completing the surveys, although depending on responses participants may have been routed to the RCT that provided compensation. This study was approved by the Institutional Review Board.

Potential participants who clicked on advertisements were taken to an eligibility screener administered online. In total, 4,783 individuals clicked the advertisements and 2,932 (61.3%) consented and started the screener. Of those, 801 (27%) were ineligible for survey participation because of demographic characteristics (female or under 18 years of age), provisional eligibility for the RCT (all of the following: age 18–29 years, male sex assigned at birth and male gender identity, not in a serious monogamous relationship lasting more than 6 months, had sex with a male, had condomless anal sex in prior 6 months, and HIV negative or unknown status), or failure to complete the screener. Potential duplicates were identified based on matching on 10 demographic characteristics (e.g., age $\pm$ 1 year, zip

code, etc.). From that analysis, 53 possible duplicate pairs or triplets were identified for careful examination on additional variables (survey date and completion time, survey responses), resulting in 33 cases that were subsequently classified as duplicates and removed from analyses. The remaining 2,098 participants were routed to various surveys. In total, 2,009 men completed at least one of the study measures for the current analyses. This sample of MSM represented a wide age range and multiple demographic characteristics in order to maximize the applicability of the linking relationship and to ensure its robustness. The average age of participants was 35.3 years ($SD = 11.6$). More information on demographic characteristics is available in Table 1.

Measures

The depression syndrome scale on the ASR is composed of 14 items meant to reflect the diagnostic criteria for MDD [17]. Unlike the empirical scales on the ASR, the syndrome scales reflect diagnostic nosology rather than statistical fit. The ASR is appropriate for individuals aged 18 to 59. Items are rated on a 3-point Likert scale, where an individual rates how true the item is for them within the last 6 months. Higher scores indicate more severe depression symptoms. Previously the ASR depression scale, together with the full ASR, has been used among MSM to track mental health concerns,[21] though no studies of which we are aware have directly assessed its dimensionality among MSM or other sexual minorities.

The depression subscale on the BSI is composed of 6 items meant to screen for depression [18]. Most items relate to cognitive symptoms, though one item relates to self-harm. It is appropriate for individuals aged 13 years and older. Items are rated on a 5-point Likert scale. As with the ASR, higher scores are indicative of greater depression, but the recall period is one week. Both the depression subscale and the BSI global severity have previously been used in research among MSM, [22; 23] though as far as we are aware, no studies have directly assessed its dimensionality in this population.

For this study, an 8-item short form from the adult PROMIS Depression item bank was used. The full bank consists of 28 items rated on a 5-point Likert scale with a one week recall period. The items can be administered via a CAT or as part of a short form. Items were developed using both qualitative and quantitative methods [2; 3; 5]. PROMIS Depression item content focuses on cognitive and emotional manifestations of depression over somatic symptoms such as changes in diet, energy, or sleep. Therefore, it does not capture the full range of DSM-defined diagnostic criteria for MDD [6]. Nonetheless, the exclusion of these items eliminates potential confounding effects in some patient populations (e.g. those with comorbid physical conditions) and provides greater support to the overall statistical model. This is one of the first studies to use the PROMIS Depression domain among MSM.

Although all three measures quantify depression, item content, response options, and recall period vary. These differences were not of vital importance for the RADAR study. Rather, the goal was to track depression as a generic mental health construct over time and to make comparisons across sub-groups. Thus, consistent with previous linking studies, the raw sum score as opposed to the proprietary scoring methods were used to map the ASR and BSI scores onto the PROMIS metric.

Study Design and Linking Methods

Linking can be accomplished using a variety of study designs and statistical approaches [24]. In this study, equipercentile and IRT-based linking approaches were compared, using a standard non-equivalent anchor test (NEAT) or a hybrid-NEAT design, respectively. The 8-item PROMIS short form was used for the calibrated anchor test, given that it well-represents the full item bank and is already in widespread use.[5] This is fewer items than was used in previous linking studies (e.g. the CES-D was linked with the full item bank, the PHQ-9 had a 20-item anchor, and the BDI-II had a 15-item anchor).[10] The hybrid-NEAT design is also referred to as *common item equating* [or linking] *to a calibrated pool*. [24] In these cases, a subset of items from an IRT-calibrated bank are used to link candidate items to that metric. The items may both compose a legitimate full test (i.e. a complete short form) and be a subset of a larger instrument (i.e. an anchor).

The equipercentile approach matches percentile ranks on the score distributions to create a nonlinear relationship between the measures. This transformation tends to result in a jagged pattern that is sensitive to random sampling error, sample idiosyncrasies, and overall sample size. As such, smoothing the score distribution is recommended to improve the linking relationship [24]. For this study, the percentile ranks for the raw sum score on the ASR and BSI were matched to the scaled PROMIS T-score (i.e. direct raw-to-scale equipercentile linking).

IRT-based linking approaches maximize the calibrated parameters by either fixing the anchor parameters to their known values (*fixed-anchor calibration*), or by estimating all parameters and using linking coefficients computed using the anchor items to convert the parameters from all items back to the original metric (*linked separate calibration*; LSC). Both methods were considered for this study. For the LSC method, Stocking-Lord [25] linking coefficients were calculated, which minimize the differences in the test characteristic curves (TCC) for the anchor items. Consistent with the PROMIS Depression item bank, all items were fit to the graded response model [26].

Linking Assumptions and Statistical Analyses

There are five assumptions relevant to linking and equating studies: the scales should measure the same construct; they should have equal reliability; linking should be bi-directional or symmetrical; scoring results should not depend on the test given; and links should be population invariant [27]. Each of these assumptions was considered prior to final item calibration or to evaluate the linking relationship.

First is the *same construct requirement*. Several approaches were used to evaluate this including correlating the mean item scores on each scale, evaluating unidimensional factor models on the aggregated item set, and examining potential multidimensional aggregated item structures. Unidimensional IRT on the aggregated items is statistically equivalent to confirmatory factor analysis [28]. Unrestricted and restricted multidimensional bifactor models were also considered [29]. Bifactor modeling also allows the ability to identify and model locally dependent item subsets.[30] Unrestricted models were estimated using the Comprehensive Exploratory Factor Analysis (CEFA) program, using raw data to calculate

the polychoric correlation matrix.[31] The extracted factor loadings were rotated to a bi-quartimin rotation using the GPArotation package in the R statistical environment.[32; 33] Restricted unidimensional and multidimensional models were fit using flexMIRT3,[34] using the EM algorithm (with dimension reduction for bifactor models) [30] for estimation. Several fit statistics were used to compare models, including the Root Mean Square Error of Approximation (RMSEA), the Tucker Lewis Index (TLI), and Akaike Information Criterion (AIC). RMSEA values less than 0.05 are generally considered “good” and values greater than 0.10 are considered “poor” [35]. TLI values greater than 0.90 are considered acceptable [36]. These cutoffs, however, were developed within a factor analysis framework, and their utility within IRT is less clear [37]. AIC values are data-specific and as such do not have set cutoffs; rather the lower value is to be preferred. For seemingly multivariate items, the Item Explained Common Variance (I-ECV) statistic was calculated [38]. I-ECV values range from 0 to 1, with lower values suggesting a greater departure from the *same construct requirement*.

The second assumption is the *equal reliability requirement*. This assumption is often violated, primarily due to using an anchor across tests, and yet produces satisfactory linking relationships [27]. Test anchors are often shorter than the full test, as is the case in this study, where an 8-item short form is used instead of the full PROMIS Depression item bank. The short form is broadly comparable to the full bank,[5] but like most anchor-based linking studies, violates this assumption to a small extent. However, in an IRT framework, this assumption may also be less relevant, as the test information function adjusts for unequal reliabilities at an item-level. Nonetheless, Cronbach’s alpha was calculated as a measure of internal consistency reliability for the individual and aggregated scales.

The *symmetry requirement* is third, which states that scores from each measure should be transformable onto the other scale. Both the equipercentile and IRT-based methods utilized in this study conform to this requirement.

The *equity requirement* states that scoring should not vary depending on the test given. In this study, linking the ASR and BSI to the PROMIS metric creates three types of scores. The first type (three similar but unique scores per measure) were linked from the ASR or BSI raw sum score using equipercentile methods with none, less, or more smoothing. The second type of score (one score per measure) used the ASR and BSI item parameters to create crosswalk *expected a posteriori* (EAP) scores. The third type of score was the pattern-based EAP score from the PROMIS short form. The linked scores should highly correlate with and well-approximate the obtained PROMIS score. The difference between actual PROMIS scores and linked scores was also calculated and summarized, including the Root Mean Square Deviation (RMSD).

The *population invariance requirement* states that a linking relationship should be equally applicable to all subgroups. Traditionally this involves examining subgroup-specific linking relationships to an overall population-based link [27]. In this study, a novel approach was taken to evaluate population invariance. Items were evaluated for differential item functioning (DIF) across subgroups. If items exhibited DIF, subgroup specific linking parameters may have been warranted. This approach had been suggested [8], but previous

analyses have generally avoided population invariance evaluations or have used subgroup specific linking functions and calculated the standardized root expected mean square difference. The impact of potential violations to this requirement were then evaluated using the weighted area between the expected score curve (wABC) [39]. Values greater than 0.30 indicate scoring likely would be affected for the subgroups. Meaningful subgroups considered herein include age (dichotomized above and below 25 years), HIV status (negative, positive, or unknown), and race (Caucasian-only and non-Caucasian or biracial).

A primary threat to the population invariance requirement, however, relates to gender or sexual minority status. This sample does not allow direct generalization to the broader population. However, IRT-calibrated parameters should be invariant across samples, and we are assuming that invariance for the PROMIS items. Under that assumption, mean depression severity may differ in this sample, which would be captured by the estimated score distribution. The ASR and BSI item calibrations will be on the PROMIS general population metric as well. With reasonable caution, we suggest the results of this study can be generalized under this strong assumption unless or until future research demonstrates a violation in the population invariance requirement due to gender or sexual minority status.

Results

All three scales were designed to measure the same construct: depression. However, examination of item content suggested that the one item on the BSI and two items on the ASR also measured self-harm, and several items on the ASR captured somatic manifestations of depression, consistent with the DSM-5 nosology but not included in the PROMIS item bank. The mean item scores correlated highly between PROMIS and the ASR ($r = .82$) and BSI ($r = .89$). This suggested that it was appropriate to proceed with a linking analysis.

Unrestricted and restricted item factor analysis also supported the *same construct requirement*. A unidimensional model fit the data relatively well (RMSEA = .06; TLI = 1.00; AIC = 85385.47). Nonetheless, unrestricted multidimensional models were considered. Initial unrestricted models suggested a bifactor structure with specific factors due to local dependence among the three self-harm and five somatic symptom items. This minimally affected fit when submitted to a restricted item factor analysis (RMSEA = .06; TLI = 1.00; AIC = 83927.03). I-ECV values across the items reflecting these specific factors ranged from 0.46 to 0.74.

The internal consistency of each scale individually and when aggregated with PROMIS was calculated to evaluate the *equal reliability requirement* (see Table 2). Examination of Cronbach's alpha and the minimum, mean, and maximum corrected item-total correlation demonstrated that PROMIS had a better internal consistency than the other scales, but all were well within an appropriate range, especially when aggregated.

Both the IRT and equipercentile linking methods meet the *symmetry* requirement for linking. Thus the *equity* requirement is the primary means by which to choose between linking methods. To create the equipercentile linking relationships, the raw score distribution

for the ASR and BSI were separately compared to the PROMIS T-score distribution. The summed scores were indexed from 0; thus ranged from 0 to 28 on the ASR and from 0 to 24 on the BSI. The relationship was evaluated with none, less (0.3) and more (1.0) smoothing. For IRT-based linking, items from all three scales were aggregated and fit to a unidimensional model. Item parameters were estimated twice: fixed-anchor, and Stocking-Lord LSC. In the former case, the estimated sample mean and SD on the PROMIS metric were 54.4 and 9.6, respectively. In the latter case, Stocking-Lord coefficients (multiplicative = 0.94, additive = 0.44) were calculated to place the items on the original PROMIS metric. Differences between the TCCs for fixed-anchor and LSC were less than 0.25 units across the T-score range. For this reason, the fixed-anchor parameters were used (See Table 3) to create an IRT-based sum-score crosswalk table [40; 41].

The *equity* requirement states that a score should not vary by the test given after linking. We extend this concept to state that the optimal linking method should be the one which best recovers the criterion: the obtained PROMIS Depression score. All available data within the sample was used to correlate scores and calculate mean differences. Table 4 presents full-sample correlations, mean differences (obtained minus linked score), standard deviations of differences, and the RMSD between linked and actual PROMIS Depression T-scores. The linked scores were broadly comparable regardless of linking method. However, the minimized RMSD suggested the IRT-based links were marginally superior to the equipercentile methods. Across methods, the mean difference, which is a measure of linking bias, was less than one point on the T-score metric. Given these results, the fixed-anchor IRT-based crosswalk (Table 5) was chosen as the optimal method to map ASR and BSI raw sum scores onto the PROMIS metric. Note that this MSM population had a different distribution than the general population, as described above. However, in order to maximize the applicability of the results from this study to other populations, the general population distribution was used when developing the sum score conversions.

Overall, the linking relationship was better for the BSI than the ASR. The linked scores from the BSI have a smaller mean difference, and offer greater precision. The EAP from sum score linked scores have a marginal reliability of 0.82 and 0.80 for the BSI and ASR, respectively.

Finally, linking relationships should exhibit *subpopulation invariance*. To ensure this, DIF analyses were considered on each of the three scales. All parameters were freely estimated in these analyses as the focus was population invariance, not measurement in relationship to an existing metric. The two-stage Wald statistic flagged several items for potential DIF (three PROMIS, three ASR, and two BSI items for age-related DIF; one PROMIS and one ASR item for HIV-status-related DIF; no items for race-related DIF). Follow-up analyses suggested all of these differences minimally affected scoring (all wABC values < 0.20) and thus were of negligible significance. *Population invariance* is expected to hold across the linking relationship.

Discussion

The PROMIS metric has been proposed as a common scoring method for reporting depression scores. [8; 10] This study extends the common metric by linking the ASR and BSI to it. There are several strengths in this study. Common item linking to a calibrated item pool with a hybrid-NEAT design places new items on an existing metric in a robust manner without needing to administer all items to all participants [24]. The ability to crosswalk between scores on multiple measures allows researchers and clinicians to compare outcomes across studies and over time in ways that were not as accessible previously.

The ASR and BSI met the statistical requirements to proceed with linking. They were broadly unidimensional, had similar reliabilities, and comparable content across measures. There were some differences, however, related to content measured by these scales that is not native to the PROMIS measures: self-harm and somatic symptoms. These locally-dependent items, within and across legacy measures, was not severe enough to prevent linking the measures.

Linking proceeded using equipercentile and IRT-based methods. Both of these ensure a symmetric linking relationship. However, in order to pick the optimal method, we relied on the most equitable for score recovery. The mean difference (bias) between linked and actual scores was small across all methods, but the fixed anchor IRT-based method had the smallest RMSD, suggesting its marginal superiority. IRT-based methods of assessing subpopulation invariance also were considered, to ensure that the linking was robust across subject subgroups.

Implications

These results have direct implications for ongoing research. They allow analysts to merge multiple longitudinal samples of MSM in order to examine depression symptoms in larger and more diverse samples. The RADAR study, for which these analyses were conducted merges two origination samples (Project Q2 and Crew 450) together with a new sample into one larger longitudinal cohort of young MSM. Existing data used different depression questionnaires. Aggregating data collected at prior waves with prospective data collected benefits the current study. It will allow us to examine how depression symptoms change over a larger age range, to analyze cohort differences in longitudinal trajectories of depressive symptoms, and to conduct these analyses in a larger sample.

This unified metric has multiple novel research and clinical applications. Applications of linking health measures have not been well defined, but we see a role in several arenas, including longitudinal research, improved meta-analyses, and quality improvement efforts. Multiple national and international quality standards for depression exist. Converting to a unified metric would allow clinics, hospitals, and international collaborations utilizing various depression measures to ensure adequate quality and ongoing improvement in mental health care provision.

Reasonable Cautions

Although the results from this study were robust, caution needs to be exercised for two reasons: first, the sample was not representative of either the USA general population or a clinical depression sample, and second, linked scores have the potential for increased imprecision. Hybrid- and traditional NEAT designs allow for non-equivalent groups under the assumption that the anchor items are invariant in the current sample.[24] However, utilizing an exclusively MSM sample has the potential that the relationship violates population invariance in untestable ways (e.g. gender differences). Linking places the new item parameters on the same scale as the anchor items, and thus provided the anchor items are invariant, the legacy measures should be equally applicable for the general population. This generalization is built on the strong assumption related to invariance of the anchor parameters and invariance of the linking relationship. This caution, while possible, risks *reductio ad absurdum*. Unless bias or DIF can be empirically demonstrated at a later date, linking scores can be extrapolated to other subgroups and the broader population with reasonable caution. It is more likely that depression is different in terms of average severity, as was evident in the distribution of scores in this sample, than in structural form. Extrapolating the linking relationship with caution is consistent with previous recommendations in unique-sample linking studies [8].

Increased error is a second reasonable concern in a linking study. There is error associated with unreliability in the original instruments and bias and error in the linking relationship. Thus switching between measures may affect reliability, especially on an individual basis. Reliability increases for tracking the mean score of a group over time or comparing the mean score across groups on different measures. While the RMSD was minimized for IRT-based linking, it remained rather wide for estimating individual scores. Transforming group-level scores should be recommended, as it further minimizes the error associated with linking.

Researchers and clinicians should be aware of these concerns; however, insofar as the ASR and BSI parameters were placed on the PROMIS metric, the use of individual linked scores are the best estimate of PROMIS Depression without using a PROMIS measure. Future studies should examine whether our assumption of subpopulation invariance in the anchor and linking relationship is supported. Unless invariance is empirically demonstrated, this linking relationship allows clinicians and researchers to rescore the ASR and BSI onto the common metric, shared by several other common depression questionnaires.[10]

Acknowledgments

Funding: This study was supported by the National Institute on Drug Abuse (U01DA036939). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

References

1. Kessler RC, Berglund P, Demler O. The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (NCS-R). JAMA. 2003; 289(23):3095–3105. [PubMed: 12813115]

2. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, Devellis R, Dewalt D, Fries JF, Gershon R, Hahn EA, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R, Lai JS. PROMIS Cooperative Group. 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*. 2010; 63(11): 1179–1194. [PubMed: 20685078]
3. DeWalt DA, Rothrock N, Yount S, Stone AA. PROMIS Cooperative Group. Evaluation of Item Candidates: The PROMIS Qualitative Item Review. *Medical Care*. 2007; 45(5 Suppl 1):S12–S21. [PubMed: 17443114]
4. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clinical & Experimental Rheumatology*. 2005; 23(5 Suppl 39):S53–57.
5. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): depression, anxiety, and anger. *Assessment*. 2011; 18(3):263–283. [PubMed: 21697139]
6. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. 5. Washington, DC: American Psychiatric Association; 2013.
7. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, Hays RD. Representativeness of the PROMIS Internet Panel. *Journal of Clinical Epidemiology*. 2010; 63(11):1169–1178. [PubMed: 20688473]
8. Gibbons LE, Feldman BJ, Crane HM, Mugavero M, Willig JH, Patrick D, Schumacher J, Saag M, Kitahata MM, Crane PK. Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. *Quality of Life Research*. 2011; 20(9): 1349–1357. [PubMed: 21409516]
9. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9 Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*. 2001; 16(9):606–613. [PubMed: 11556941]
10. Choi SW, Schalet B, Cook KF, Cella D. Establishing a Common Metric for Depressive Symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. *Psychological Assessment*. 2014; 26(2):513–527. [PubMed: 24548149]
11. Beck, AT., Steer, RA., Brown, GK. *Manual for the Beck Depression Inventory–II*. San Antonio, TX: Psychological Corporation; 1996.
12. Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*. 1977; 1(3):385–401.
13. Reeve B, Thissen D, DeWalt D, Huang IC, Liu Y, Magnus B, Quinn H, Gross H, Kisala P, Ni P, Haley S, Mulcahey MJ, Charlifue S, Hanks AR, Slavin M, Jette A, Tulskey D. Linkage between the PROMIS® pediatric and adult emotional distress measures. *Quality of Life Research*. 2016; 25(4): 823–833. [PubMed: 26424169]
14. Olino TM, Yu L, McMakin DL, Forbes EE, Seeley JR, Lewinsohn PM, Pilkonis PA. Comparisons Across Depression Assessment Instruments in Adolescence and Young Adulthood: An Item Response Theory Study Using Two Linking Methods. *Journal of Abnormal Child Psychology*. 2013; 41(8):1267–1277. [PubMed: 23686132]
15. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An Inventory for Measuring Depression. *Archives of General Psychiatry*. 1961; 4:561–571. [PubMed: 13688369]
16. Wahl I, Löwe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, Aita SA, Bergemann N, Brähler E, Rose M. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. *Journal of Clinical Epidemiology*. 2014; 67(1):73–86. [PubMed: 24262771]
17. Rescorla, LA., Achenbach, TM. The Achenbach System of Empirically Based Assessment (ASEBA) for Ages 18 to 90 Years. In: Maruish, ME., editor. *The use of psychological testing for treatment planning and outcomes assessment*. 3. Mahwah, NJ: Lawrence Erlbaum Associates; 2004. p. 115–152.
18. Derogatis, LR. *The Brief Symptom Inventory: administration, scoring & procedures manual*. Minneapolis, MN: National Computer Systems; 1993.

19. Mustanski BS, Garofalo R, Emerson EM. Mental health disorders, psychological distress, and suicidality in a diverse sample of lesbian, gay, bisexual, and transgender youths. *American journal of public health*. 2010; 100(12):2426–2432. [PubMed: 20966378]
20. Garofalo R, Hotton AL, Kuhns LM, Gratzner B, Mustanski B. Incidence of HIV infection and Sexually Transmitted Infections and Related Risk Factors among Very Young Men Who Have Sex with Men. *Journal of Acquired Immune Deficiency Syndromes*. 2016; 72(1):79–86. [PubMed: 26745827]
21. Puckett JA, Newcomb ME, Garofalo R, Mustanski B. The Impact of Victimization and Neuroticism on Mental Health in Young Men Who Have Sex with Men: Internalized Homophobia as an Underlying Mechanism. *Sexuality Research and Social Policy*. 2016; 13(193):1–9. [PubMed: 26858776]
22. Mustanski B, Garofalo R, Herrick A, Donenberg G. Psychosocial health problems increase risk for HIV among urban young men who have sex with men: preliminary evidence of a syndrome in need of attention. *Annals of Behavioral Medicine*. 2007; 34(1):37–45. [PubMed: 17688395]
23. Dowshen N, Binns HJ, Garofalo R. Experiences of HIV-related stigma among young men who have sex with men. *AIDS Patient Care and STDs*. 2009; 23(5):371–376. [PubMed: 19320600]
24. Kolen, MJ., Brennan, RL. Test equating, scaling, and linking : methods and practices. 3. New York, NY: Springer; 2014.
25. Stocking ML, Lord FM. Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*. 1983; 7(2):201–210.
26. Samejima, F. The general graded response model. In: Nering, ML., Ostini, R., editors. *Handbook of polytomous item response theory models*. New York, NY: Routledge; 2010. p. 77-107.
27. Dorans NJ, Holland PW. Population Invariance and the Equatability of Tests: Basic Theory and The Linear Case. *Journal of Educational Measurement*. 2000; 37(4):281–306.
28. Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. *Psychological Methods*. 2007; 12(1):58–79. [PubMed: 17402812]
29. Reise SP. The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*. 2012; 47(5):667–696. [PubMed: 24049214]
30. Cai L. A two-tier full-information item factor analysis model with applications. *Psychometrika*. 2010; 75(4):581–612.
31. Browne, MW., Cudeck, R., Tateneni, K., Mels, G. CEFA: Comprehensive Exploratory Factor Analysis. Columbus, OH: Ohio State University; 2010.
32. Bernaards CA, Jennrich RI. Gradient Projection Algorithms and Software for Arbitrary Rotation Criteria in Factor Analysis. *Educational and Psychological Measurement*. 2005; 65(5):676–696.
33. Jennrich RI, Bentler PM. Exploratory bi-factor analysis: The oblique case. *Psychometrika*. 2012; 77(3):442–454. [PubMed: 27519775]
34. Houts, CR., Cai, L. flexMIRT® : Flexible Multilevel Multidimensional Item Analysis and Test Scoring (Version 3.0). Chapel Hill, NC: Vector Psychometric Group; 2015.
35. Browne MW, Cudeck R. Alternative ways of addressing model fit. *Sociological Methods and Research*. 1992; (21):230–258.
36. Hu, L., Bentler, PM. *Psychological Methods*. 1998. Fit indices in covariance structure modeling: Sensitivity to underparameterization model misspecification; p. 424-453.
37. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*. 2009; 18(4):447–460. [PubMed: 19294529]
38. Stucky BD, Thissen DM, Edelen MO. Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement*. 2012; 37(1):41–57.
39. Edelen M, Stucky B, Chandra A. Quantifying 'problematic' DIF within an IRT framework: application to a cancer stigma index. *Quality of Life Research*. 2015; 24(1):95–103. [PubMed: 24214177]
40. Orlando M, Sherbourne CD, Thissen D. Summed-score linking using Item Response Theory: Application to depression measurement. *Psychological Assessment*. 2000; 12(3):354–359. [PubMed: 11021160]

41. Thissen D, Pommerich M, Billeaud K, Williams VSL. Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses. *Applied Psychological Measurement*. 1995; 19(1):39–49.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Demographic Characteristics

		N	%
Orientation	Gay	1691	84.2%
	Bisexual	228	11.3%
	Queer	37	1.8%
	Questioning or Unsure	26	1.3%
	Heterosexual	7	0.3%
	Other	17	0.8%
	[No Response]	3	0.1%
Race & Ethnicity	White/Caucasian	1220	60.7%
	Black/African-American	160	8.0%
	Latino/Latina	365	18.2%
	American Indian	16	0.8%
	Asian	87	4.3%
	Pacific Islander	5	0.2%
	Other	55	2.7%
	[No Response]	101	5.0%
HIV Status	Negative	1428	71.1%
	Positive	310	15.4%
	Unknown	269	13.4%
	[No Response]	2	< 0.1%

Table 2

Classical Reliability Information

Scale(s)	Number of Items	Cronbach's Alpha	Corrected Item-Total Correlation		
			Minimum	Average	Maximum
ASR	14	.89	.35	.58	.75
BSI	6	.91	.61	.76	.85
PROMIS	8	.96	.81	.85	.90
ASR+PROMIS	22	.95	.34	.67	.86
BSI+PROMIS	14	.97	.61	.81	.89
ASR+BSI+PROMIS	28	.97	.34	.69	.87

Note: All correlations are significant at the $p < .01$ level. No scale would have had a higher alpha if an item were dropped.
Abbreviations: ASR=Adult Self Report; BSI=Brief Symptom Inventory; PROMIS=Patient Reported Outcomes Measurement Information System

Table 3

Fixed Anchor Calibration Item Parameters

Item Number ^a	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
BSI 1	2.39	0.23	1.17	1.89	2.82
BSI 2	2.23	-0.39	0.58	1.19	2.04
BSI 3	3.10	-0.07	0.87	1.48	2.32
BSI 4	4.37	0.54	1.16	1.62	2.22
BSI 5	3.46	0.30	1.05	1.49	2.03
BSI 6	2.44	1.39	2.07	2.48	3.16
ASR 1	1.23	1.36	3.31		
ASR 2	1.65	2.79	4.20		
ASR 3	0.83	-0.62	2.02		
ASR 4	2.96	0.53	1.87		
ASR 5	1.76	0.81	2.26		
ASR 6	1.69	0.25	1.70		
ASR 7	2.32	0.94	2.16		
ASR 8	1.10	1.03	2.57		
ASR 9	1.35	0.58	2.32		
ASR 10	1.84	1.60	3.00		
ASR 11	1.06	0.13	2.16		
ASR 12	1.72	0.17	1.98		
ASR 13	3.35	0.26	1.60		
ASR 14	2.88	0.59	1.78		

Note: The item parameters are on the theta-metric (mean 0, SD 1). EAP scores are calculated and then transformed to the T-score metric by multiplying by 10 and adding 50 for the score and by multiplying by 10 for the SD.

Abbreviations: ASR=Adult Self Report; BSI=Brief Symptom Inventory

^aItem numbers here do not correspond with item numbers on the originating scales

Table 4

Full-Sample Comparisons between Linked and Actual Scores

Scale	Linking Method	Correlation	Mean Difference	SD of Differences	RMSD
ASR	Equipercntile None	.79	0.30	6.35	6.35
	Equipercntile Less	.78	0.72	6.94	6.98
	Equipercntile More	.78	0.76	7.04	7.08
	IRT-Fixed Anchor Crosswalk	.80	0.64	6.29	6.32
BSI	Equipercntile None	.86	0.17	5.21	5.21
	Equipercntile Less	.86	0.11	5.23	5.23
	Equipercntile More	.85	0.35	5.46	5.47
	IRT-Fixed Anchor Crosswalk	.87	0.35	5.09	5.10

Abbreviations: ASR=Adult Self Report; BSI=Brief Symptom Inventory; IRT=Item Response Theory; RMSD=Root Mean Square Difference

Table 5

IRT-Based Sum Score Crosswalk to the Unified Depression Metric

Sum Score	ASR to PROMIS T-Score	ASR to PROMIS SD	BSI to PROMIS T-Score	BSI to PROMIS SD	Sum Score
0	36.5	6.5	38.7	6.1	0
1	40.6	5.7	45.0	4.3	1
2	43.7	5.2	48.3	3.8	2
3	46.5	4.6	50.6	3.5	3
4	48.9	4.1	52.6	3.3	4
5	51.0	3.8	54.3	3.0	5
6	52.8	3.5	55.9	2.9	6
7	54.4	3.3	57.3	2.8	7
8	56.0	3.2	58.6	2.7	8
9	57.4	3.2	59.8	2.7	9
10	58.7	3.1	61.0	2.6	10
11	60.0	3.1	62.1	2.6	11
12	61.2	3.1	63.2	2.6	12
13	62.5	3.0	64.4	2.6	13
14	63.7	3.0	65.5	2.6	14
15	65.0	3.0	66.6	2.6	15
16	66.3	3.0	67.7	2.6	16
17	67.5	3.0	68.9	2.7	17
18	68.8	3.0	70.1	2.7	18
19	70.1	3.1	71.4	2.8	19
20	71.4	3.2	72.8	3.0	20
21	72.8	3.3	74.3	3.1	21
22	74.2	3.4	75.9	3.2	22
23	75.7	3.5	78.1	3.5	23
24	77.4	3.7	81.3	4.3	24
25	79.2	3.9			25
26	81.2	4.2			26
27	83.5	4.5			27

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Sum Score	ASR to PROMIS T-Score	ASR to PROMIS SD	BSI to PROMIS T-Score	BSI to PROMIS SD	Sum Score
28	86.3	4.9			28

Abbreviations: ASR=Adult Self Report; BSI=Brief Symptom Inventory; PROMIS=Patient Reported Outcomes Measurement Information System

Note: This crosswalk table assumed a prior distribution consistent with the PROMIS metric centered on the 2000 USA general population (i.e. mean=50.0, SD=10.0), though the MSM sample used to develop the relationship had a slightly different distribution (mean=54.4, SD=9.6)