Published in final edited form as:
Qual Life Res. 2011 November ; 20(9): 1349–1357. doi:10.1007/s11136-011-9882-y.

Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures

Laura E. Gibbons,
General Internal Medicine, University of Washington, Box 359780, Harborview Medical Center, 325 Ninth Ave, Seattle, WA 98104, USA, gibbonsl@u.washington.edu

Betsy J. Feldman,
Allergy and Infectious Diseases, University of Washington, Box 359931, Harborview Medical Center, 325 Ninth Ave, Seattle, WA 98104, USA

Heidi M. Crane,
Allergy and Infectious Diseases, University of Washington, Box 359931, Harborview Medical Center, 325 Ninth Ave, Seattle, WA 98104, USA

Michael Mugavero,
Department of Medicine, Division of Infectious Disease, University of Alabama at Birmingham, 1530 Third Ave S, CCB 142, Birmingham, AL 35294-2050, USA

James H. Willig,
Department of Medicine, Division of Infectious Disease, University of Alabama at Birmingham, 1530 Third Ave S, CCB 178, Birmingham, AL 35294-2050, USA

Donald Patrick,
Department of Health Services, University of Washington, Box 359455, 4333 Brooklyn Ave NE, Seattle, WA 98195-9455, USA

Joseph Schumacher,
Division of Preventive Medicine, School of Medicine, University of Alabama at Birmingham, 1717 11th Avenue South, 616 Medical Towers Building, Birmingham, AL 35209, USA

Michael Saag,
Center for AIDS Research, University of Alabama at Birmingham, 845 19th Street South, BBRB 256, Birmingham, AL 35294-2050, USA

Mari M. Kitahata, and
Allergy and Infectious Diseases, University of Washington, Box 359931, Harborview Medical Center, 325 Ninth Ave, Seattle, WA 98104, USA

Paul K. Crane
General Internal Medicine, University of Washington, Box 359780, Harborview Medical Center, 325 Ninth Ave, Seattle, WA 98104, USA

Abstract

Correspondence to: Laura E. Gibbons.

Electronic supplementary material The online version of this article (doi:10.1007/s11136-011-9882-y) contains supplementary material, which is available to authorized users.

Purpose—We provide detailed instructions for analyzing patient-reported outcome (PRO) data collected with an existing (legacy) instrument so that scores can be calibrated to the PRO Measurement Information System (PROMIS) metric. This calibration facilitates migration to computerized adaptive test (CAT) PROMIS data collection, while facilitating research using historical legacy data alongside new PROMIS data.

Methods—A cross-sectional convenience sample (n = 2,178) from the Universities of Washington and Alabama at Birmingham HIV clinics completed the PROMIS short form and Patient Health Questionnaire (PHQ-9) depression symptom measures between August 2008 and December 2009. We calibrated the tests using item response theory. We compared measurement precision of the PHQ-9, the PROMIS short form, and simulated PROMIS CAT.

Results—Dimensionality analyses confirmed the PHQ-9 could be calibrated to the PROMIS metric. We provide code used to score the PHQ-9 on the PROMIS metric. The mean standard errors of measurement were 0.49 for the PHQ-9, 0.35 for the PROMIS short form, and 0.37, 0.28, and 0.27 for 3-, 8-, and 9-item-simulated CATs.

Conclusions—The strategy described here facilitated migration from a fixed-format legacy scale to PROMIS CAT administration and may be useful in other settings.

Keywords

Calibration; Computerized adaptive testing; Depression; Item banks; Item response theory; PROMIS

Introduction

The Patient-Reported Outcomes (PRO) Measurement Information System (PROMIS) initiative has developed scales for many health-related constructs, including physical functioning, fatigue, and emotional distress [1-3]. Test items are designed to measure each construct's entire severity continuum and can be administered using computer adaptive testing (CAT). PROMIS items were calibrated to a normative sample representing the general US population [2].

CAT uses a respondent's prior item responses to determine which item from an item bank to administer next or whether measurement is precise enough to terminate further data collection [4-7]. CAT almost always offers improved measurement precision for a given number of items, as compared with fixed-format administration [5-8]. This greater precision means fewer items may be needed to achieve the same quality of measurement, reducing patient burden.

Given the advantages of using PROMIS measures and CAT administration, many clinicians and researchers may want to assess PRO domains using PROMIS CAT. However, practical concerns may arise in migrating to PROMIS CAT from using another instrument for the same domain, which we call the "legacy" instrument. Clinicians may be used to interpreting scores on the metric of the legacy instrument, and the display of scores using both the legacy and PROMIS metrics may be needed to facilitate comfort with the PROMIS metric. Data may have been collected over many years using the legacy instrument, and researchers may want to continue to use historical data. As time goes on and experience accrues with a legacy instrument, it may become increasingly difficult to justify switching to a new metric, however appealing it may be, unless there is a way to retain historical legacy information.

In this paper, we show how to overcome these obstacles to migrating to PROMIS CAT. We demonstrate tools to facilitate equating legacy instrument scores to PROMIS scores. This work expands on traditional score linking methods (see [9]) and our prior work calibrating

tests using a common items design [10]. Here we have calibrated two depression symptom scales using a single group design, administering the legacy instrument, the Patient Health Questionnaire (PHQ-9), and a subset of PROMIS items (the PROMIS Depression short form) to the same group of patients. This work expands on an earlier series of papers by Bjorner et al. that demonstrated migration to a new headache measure using similar strategies [4, 11, 12]. Here we show how to incorporate PROMIS item bank parameters that are treated as known and fixed to facilitate migration from a legacy instrument to PROMIS CAT. We provide the syntax needed to accomplish the analyses demonstrated here, written in readily modifiable files.

Methods

An overview of the migration process is given in Table 1.

Participants

Participants were a cross-sectional convenience sample from the Universities of Washington (UW) and Alabama at Birmingham (UAB) HIV clinics, both Centers for AIDS Research Network of Integrated Clinical Systems clinical sites [13] (Table 2). All participants were in routine clinical care and completed PRO assessments between August 2008 and December 2009.

HIV-infected patients over 18 years of age completed a multi-domain assessment in clinics prior to routinely-scheduled appointments, using open-source web-based survey software on touch-screen PCs connected to a wireless network [14, 15]. The first assessment at which the PROMIS depression short form was administered was included in this analysis. Patients unable to provide informed consent, such as those with dementia, or patients who did not speak English or Spanish did not participate in the survey. The institutional review boards from both sites approved the study, which was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. All participants gave their informed consent prior to completing the assessment.

Measures

Step 1—We collected depression symptom data from two instruments: the 8-item PROMIS depression short form [2, 8, 16] and the 9-item PHQ-9 depression measure [17, 18]. The PHQ-9 was designed to tap all the DSM-4 depression elements, including cognitive and somatic symptoms and activity level, while the PROMIS depression short form is focused on emotional content. The items are provided in Online Resource 1. There are no items in common, and only partial content overlap. Each scale has been subject to thorough psychometric evaluations. Each has been found to be sufficiently unidimensional to analyze using item response theory (IRT) [8, 16, 19].

Step 2a—DIF has been shown to have minimal score impact on each of these measures [16, 19, 20]. In previous analyses, we have shown that the PHQ-9 items showed little DIF with respect to a large number of covariates [19]. Here we are transforming the scale of the PHQ-9 IRT to the PROMIS metric, and DIF should not affect that transformation.

Statistical methods

Evaluation of dimensionality

Step 2b: A critical assumption of scale calibration is that both scales can be considered to measure the *same* unidimensional construct and that the PHQ-9 items can be considered to be indicators of the latent trait measured by the PROMIS depression item bank (Step 2). To

assess this assumption, we fit several confirmatory factor analysis (CFA) models using Mplus [21] code shown in Online Resource 2. We used the national item parameters from the PROMIS Version 1 item bank (http://www.nihpromis.org) for the short form items. In all analyses, we used the weighted least squares and mean-and-variance-adjusted estimator with robust standard errors [22, 23], applied to the tetrachoric correlation matrix estimated from the categorical item responses.

First, we fit a single-factor model of the PROMIS short form data using item parameters fixed to the values obtained from the item bank, which we transformed from their IRT values to values appropriate for Mplus (see Online Resource 2 for details on the transformation). The item bank parameters and our IRT analyses used the graded response model for ordinal variables [24]. We noted assessments of fit for this single-factor model including the Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), and the root mean square error of approximation (RMSEA). For PROMIS, acceptable levels of these fit statistics have been suggested: CFI > 0.95, TLI > 0.95, and RMSEA < 0.06 [2], though evidence is not available suggesting relevance for these statistics in single-factor confirmatory factor analysis models with ordinal indicators. We compared fit from the single-factor PROMIS item-only model to the fit from a single-factor model using both the PROMIS and the PHQ-9 items, again with the parameters for the PROMIS items fixed to their values from the item bank as transformed to Mplus, but with parameters for the PHQ-9 items freely estimated. We were especially interested in changes of the fit indices when we added the PHQ-9 items; small differences in fit would suggest that there was minimal impact on fit from considering the PHQ-9 items to be indicators of the latent trait measured by the PROMIS short form items. Our rationale for this approach was that prior careful analyses with large data sets have already established that the PROMIS scale can be considered sufficiently unidimensional to use IRT [2, 8, 16]. Our query then was whether adding additional indicators (the PHQ-9 items, also sufficiently unidimensional [19]) would cause a notable degradation of model fit. Negligible changes in model fit would suggest that whatever arguments could be made about the PROMIS items can be made as well about the PROMIS and PHQ-9 items.

The derivation of the PROMIS item bank considered the issue of local dependence very carefully. For the purposes of our calibration, the PROMIS items were treated as anchors, with item parameters treated as known (at their PROMIS values) and fixed for our analyses. For the PHQ-9 items, there is still the possibility that parameters could be affected by violations of local independence. We performed a sensitivity analysis by entering the largest residual correlation into the model and examining changes in the item parameters.

Next, we calculated the correlation between the factor scores with and without the PHQ-9. Finally, we ran a two-factor CFA where the PROMIS items were modeled as indicators of one factor, the PHQ-9 items were modeled as indicators of a different factor, and the correlation between those factors was determined. We are not aware of an established criterion, but we expected the correlation to exceed 0.9.

Calibration to the PROMIS scale—We conducted a single group calibration of the PHQ-9 and PROMIS short form using PROMIS item parameters. We outline the procedure below; complete step-by-step instructions along with sample code are provided in Online Resources 3–6. These steps used Stata [25], Parscale [26], and our *prepare* package for Stata, which can be freely downloaded by typing “ssc install prepar” at the Stata prompt. We reference specific Parscale parameter files from the Stata code in parentheses below.

Step 3: We freely estimated the PHQ-9 item parameters while fixing the PROMIS item parameters to their values from the PROMIS item bank, using the following steps. Specific syntax for doing this is shown in Online Resources 4–6.

- **(3a)** We created a Parscale parameter file (PHQPRO1.PAR) to edit, by freely estimating item parameters using the 17 items from both measures.
- **(3b)** We edited the resulting parameter file (PHQPRO1.PAR) by substituting PROMIS item bank values for the 8 PROMIS item parameters for the respective parameters freely estimated in step 3a. We saved this new parameter file (PHQPRO2.PAR).
- **(3c)** We used the parameter file from step 3b (PHQPRO2.PAR) to freely estimate parameters for the PHQ-9 items on the PROMIS metric, saving the new parameter file (PHQPRO3.PAR). This step provided estimates of depression symptom levels on the PROMIS metric based on all 17 items.

Step 4: We obtained IRT scores for just the PHQ-9 items calibrated to the PROMIS metric:

- **(4a)** We edited the parameter file generated from step 3c (PHQPRO3.PAR), removing the lines with parameters for the PROMIS items and saved the new parameter file (PHQPRO4.PAR). Note that it is important in this step to also change the number of items specified in the new parameter file (PHQPRO4.PAR) for the next step.
- **(4b)** We estimated PHQ-9 scores using the fixed parameters that were calibrated to the PROMIS metric (PHQPRO4.PAR).

Steps 3 and 4 result in estimates of depression symptom scores and their standard errors of measurement (SEM) based on responses to the PHQ-9 items, but scored on the PROMIS metric. These new PHQ-9 scores can be used alongside scores derived from PROMIS items in analyses of depression symptom levels. The PHQPRO4.PAR parameter file could be used with historical PHQ-9 item data to obtain PHQ-9 scores calibrated to the PROMIS metric. We also obtained PROMIS short form scores using PROMIS item bank parameters.

Step 5: The PHQ-9 is typically scored using sum (total) scores, with scores ranging form 0–27. Clinically relevant labels based on these sum scores have been promulgated for this scale [27]. We converted the PHQ-9 IRT scores to a mean of 50 (SD 10), as is PROMIS convention. A test characteristic curve (TCC), an important figure generated from IRT analyses of items, can be used to indicate the most likely PHQ-9 sum score for a given PHQ-9 or PROMIS IRT score. This will help clinicians familiar with the PHQ-9 to interpret new PROMIS-based depression symptom scores.

Assessment of measurement precision

Step 6: With the data we collected for this study, we determined the distribution of the SEMs around depression symptom levels estimated from the PROMIS short form and the PHQ-9 as produced by Parscale. Next, we simulated CAT administration for our participant cohort using the parameters from all 28 depression items in the PROMIS item bank [2]. We used Firestar, an interface that produces R code to run a CAT simulation [28]. We estimated depression levels on the PROMIS metric from the combined PROMIS depression short form and the PHQ-9 items. These depression levels were then starting points for Firestar simulations, which simulated anticipated responses to each of the 28 PROMIS items based on these estimated depression levels observed in our sample. From each CAT simulation from Firestar, we obtained the SEM around the estimated depression symptom scores. We compared several different CAT algorithms. In the first two, everyone was administered the

number of items of the PHQ-9 and PROMIS short form, 9 and 8 items, respectively. Then we simulated shorter tests to determine the minimum number of items required to surpass the measurement precision of the PHQ-9.

Results

Step 2b in calibrating the PHQ-9 to the PROMIS metric was to determine whether the PHQ-9 items could be considered to be indicators of the latent construct defined by the PROMIS short form depression items. The single-factor model for the 8 PROMIS items, with parameters fixed to the values obtained from the PROMIS item bank, had a CFI of 0.996, a TLI of 0.998 and an RMSEA of 0.154. This suggests excellent fit by the CFI and TLI, but a suboptimal fit according to the RMSEA. The fit for the single-factor model for all 17 PROMIS and PHQ-9 items was nearly identical to the PROMIS-alone model, with CFI 0.987, TLI 0.992, and RMSEA 0.160. The correlation between the PROMIS short form and the combined PROMIS-metric 17-item factor scores was 0.98 and a scatter plot of the scores is available in Online Resource 7. Finally, in the two-factor model, the correlation between the PROMIS factor and the PHQ-9 factor was 0.91. In our sensitivity analysis of the effects of local dependence, entering the largest residual correlation reduced the standardized loadings of those two items by 0.036 or less (less than 5%). These analyses suggest that the PHQ-9 items can be considered to be indicators of the same construct measured by the PROMIS items and that treating the PHQ-9 items as indicators of that single factor defined by the PROMIS short form items is as appropriate for these data as treating the PROMIS short form items as the sole indicators of that factor.

We calibrated the PHQ-9 items to the PROMIS metric using steps 3 and 4 above (Online Resources 3–6) and obtained PROMIS short form scores on the PROMIS metric as well. Item parameters for the PHQ-9 on the PROMIS metric are in Online Resource 1. The PHQ-9 items were less discriminating and had lower thresholds.

The test characteristic curve (Fig. 1) compares 2 ways of scoring item responses to the PHQ-9. On the *y*-axis is the traditional score, formed by summing responses to the 9 items. On the x-axis is the PHQ-9 score produced using IRT such that scores are calibrated with, and thus are directly comparable to, PROMIS IRT scores. Figure 1 shows the most likely PHQ-9 sum score corresponding to the PROMIS metric. Horizontal lines indicate the PHQ-9 cutpoints for mild (5–9), moderate (10–14), moderately severe (15–19) and severe (20–27) depression [27]. Using these categories, a PROMIS-metric score of less than 42 would correspond to no depression, 42–51 to mild depression, 52–63 to moderate, 64–72 to moderately severe, and 73 and higher-to-severe depression [27].

Finally, we ran a series of CAT simulations using the Firestar program, drawing items from the full 28-item PROMIS depression item bank. The mean depression symptom score as measured by the PROMIS short form and PHQ-9 depression items was 48.2 (SD 11.3) in the PROMIS metric. We based our CAT simulations on samples of 2,178 people with that mean and SD and varied the number of items administered in the simulated CATs. In Table 3, we show the mean SEMs estimated for the simulated samples for the 8- and 9-item CATs, alongside the mean SEMs observed for the fixed-format PHQ-9 and the PROMIS short form. The PROMIS short form had better measurement precision than the PHQ-9, as can be seen by its smaller SEM, and the 8- and 9-item-simulated CATs had better measurement precision than either fixed-format test. We determined that the minimum number of simulated CAT items needed to surpass the mean measurement precision of the PHQ-9 was 3, so we also provide the mean SEM for the simulated sample for a simulated 3-item CAT.

To see whether these differences in the mean SEM were consistent across the spectrum of depression symptoms, we plotted Lowess curves for the SEMs for the PHQ-9, the simulated 3-item PROMIS CAT, the PROMIS short form, and the simulated 8-item PROMIS CAT (Fig. 2). At all levels of depression symptoms, the PROMIS short form and the 3-item-simulated CAT each had a smaller mean SEM than the PHQ-9, and the 8-item-simulated CAT was at least as precise as either of the fixed tests.

Discussion

The PHQ-9 items can be considered indicators of the underlying factor measured by the PROMIS depression short form. We calibrated the two tests to a single metric using IRT. The analyses we performed enabled direct comparison of psychometric properties of the PHQ-9 and subsets of the PROMIS item bank such as the short form and simulated CATs. One result of the calibration was Fig. 1, which enables clinicians to roughly translate between PHQ-9 and PROMIS depression symptom scores, and the item parameters (Online Resource 1), which facilitate more precise calibration. The PROMIS scales had superior measurement precision than the PHQ-9 (Fig. 2). The tools we developed are detailed in Online Resources and can be readily modified to other settings.

There are several advantages that may be realized by migrating to PROMIS. The PROMIS scales have been extensively tested. Their validity has been assessed in a number of settings and patient populations [1-3, 20, 29, 30]. PROMIS scales have been calibrated so that comparisons can be made to the United States general population [2]. The PROMIS item banks are designed for CAT administration, and our findings add to the literature demonstrating increased precision of CAT compared to fixed-format administration [8, 29, 30]. The PROMIS items are more discriminating than PHQ-9 items, in part because they have more response categories. The PHQ-9 items are all multi-part (lots of use of the word “or”), which may also hinder their discrimination.

Our CAT simulations provide additional impetus to switch to PROMIS CAT. The PROMIS short form and 3 different length-simulated CATs provided measurement precision equal to or better than that provided by the PHQ-9. Better precision allows for shorter tests. While the PHQ-9 is associated with relatively minimal respondent burden (we have documented median completion times of around 1 min [19]), any time saved with PROMIS CAT could be applied to measuring other domains or reduce the overall respondent burden [14, 15]. Alternatively, with the same number of items as the PHQ-9, one could dramatically increase measurement precision. Not shown here but clearly feasible would be a middle strategy; a 5-item CAT, for example, would both reduce respondent burden and improve measurement precision.

A query that could be raised is why one should go to all of this trouble; why not just switch from a legacy instrument to a PROMIS instrument? In our view, there are several reasons to use our procedures. One of the strongest of these is to facilitate continued research using PRO data collected before and after the switch. We can re-compute PHQ-9 scores using the PROMIS metric, permitting us to use all our historical depression data in analyses (for example, see [31]).

A second reason to use our procedures is to facilitate clinical understandings of the new scale. Formal calibration permits helpful illustrations such as that shown in Fig. 1, enabling clinicians and researchers to understand the new scale based on an already familiar scale. Figure 1 suggests that clinical labels traditionally used with PHQ-9 sum scores correspond to roughly 10 point intervals (1SD) on the PROMIS depression metric [16]. It would be premature, however, to fix the PHQ-9 clinical labels to PROMIS depression scores. The

PHQ-9 cutpoints were selected for ease of application, rather than optimal diagnosis [27], and PHQ-9 scores may overestimate depression [32, 33]. Further research that includes diagnoses will be needed to establish clinical labels for PROMIS depression scores.

With suitable caution, calibrated PHQ item parameters (Online Resource 1) can be used in other populations to generate PROMIS scale scores from PHQ-9 item responses. A nice feature of IRT is that parameter estimates are invariant across samples (within a linear transformation) if assumptions underlying the item response models are met. These invariance properties apply within the range of overlap of trait-level distributions of the different samples. Here, we observed the full range of scores on the PHQ-9 and PROMIS depression measure. So while this study was conducted with HIV-infected patients, the calibration should be valid for other samples, unless there is DIF with respect to population group, which could be determined empirically.

To our knowledge, this is the first direct comparison of the psychometric properties of the PHQ-9 depression symptom scale to the PROMIS depression short form or simulated PROMIS CATs. The migration approach taken here facilitates an examination of psychometric properties of the tests. These results can inform our understanding of the striking difference in measurement precision between the PHQ-9 and various PROMIS scores (Fig. 2). The PROMIS item bank has more items that address moderately severe depression levels in the 60- to 70-point range than the PHQ-9, and several such items are included in the PROMIS short form. This produces PROMIS's improved measurement precision for individuals with depression levels in this range.

We have written detailed descriptions of the techniques we used to calibrate the PHQ-9 to the PROMIS metric and have included code that produced the analyses, in the hope that they will be useful to individuals performing similar analyses. The techniques outlined here could also be used to expand PROMIS item banks. The methods outlined here could be implemented with other software programs. We used Parscale because we have developed an array of Stata tools for it, but everything could have been done in other packages, such as Mplus or IRTPRO. Online Resource 2 illustrates how to convert PROMIS item parameters to values appropriate for Mplus and how to fix item parameters for analyses. Other researchers have undertaken a similar migration for a new headache measure [4, 11, 12]. Our paper adds to that literature by incorporating established item bank parameters and providing specific adaptable code. We demonstrated the specific application to depression, which may be of interest in a wide variety of clinical settings.

Several limitations of this paper should be noted. We are unaware of specific criteria for determining whether calibrated scales are sufficiently unidimensional. We were guided by theoretical considerations (i.e., both scales putatively measure depression, and similar constructs are assessed by the items in the two scales) and by the results of our analyses, which showed very little effect on fit when we added PHQ-9 items to the PROMIS depression scale. Second, at each stage in the analyses, we treated some item parameters as known and fixed, ignoring error in the estimates of the IRT parameters. Bjorner and colleagues [11] show a method for incorporating an error structure when one is available, but we did not have access to error values. Third, mean depression symptom levels in our cohort were slightly lower than those found in a nationally representative (non-HIV) population. Few participants endorsed very high levels of depression symptoms, so our confidence in the highest threshold parameter estimates for the PHQ-9 is somewhat less than it would be in a population with many more severely depressed individuals. Nevertheless, data presented here are from a very large clinical sample from two specialty medical clinics, thus representing the distribution of depression levels likely to be seen among HIV-infected patients in routine clinical care.

Steps may remain before adopting CAT for routine clinical use. PRO CAT may be considered a substantial modification to a measure and may require assessments of validity with actual (not simulated) CAT [4, 12, 34]. In addition, our CAT simulations used a traditional CAT algorithm that selects items based on their psychometric properties alone. However, there may be items such as the PHQ-9 question about suicidality that clinicians are interested in regardless of its psychometric properties. It should also be noted that we have evaluated the instruments on their psychometric properties alone; the PHQ-9 items are aligned to the diagnostic criteria for depression and may be preferable if diagnosis is the goal.

In summary, we have presented detailed methods for migrating from fixed-format legacy PRO collection to either the PROMIS short form or PROMIS CAT. Furthermore, we have shown how this may be done without losing historical data collected using the legacy instrument. We hope that the tools provided here will prove useful to others wishing to migrate from a legacy instrument to PROMIS.

Acknowledgments

This work was supported by National Institutes of Health grants U01 AR 057954, R01 MH 084759, P30 AI 27757, P30 AI 27767, R24 AI 067039, K23 MH 082641, and the Mary Fisher CARE Fund. The Patient-Reported Outcomes Measurement Information System (PROMIS) is an NIH Roadmap initiative to develop a computerized system measuring PROs in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS II was funded by cooperative agreements with a Statistical Center (Northwestern University, PI: David F. Cella, PhD, 1U54AR057951), a Technology Center (Northwestern University, PI: Richard C. Gershon, PhD, 1U54AR057943), a Network Center (American Institutes for Research, PI: Susan (San) D. Keller, PhD, 1U54AR057926) and thirteen Primary Research Sites (State University of New York, Stony Brook, PIs: Joan E. Broderick, PhD and Arthur A. Stone, PhD, 1U01AR057948; University of Washington, Seattle, PIs: Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD, 1U01AR057954; University of Washington, Seattle, PIs: Dagmar Amtmann, PhD, and Karon Cook, PhD1U01AR052171; University of North Carolina, Chapel Hill, PI: Darren A. DeWalt, MD, MPH, 2U01AR052181; Children's Hospital of Philadelphia, PI: Christopher B. Forrest, MD, PhD, 1U01AR 057956; Stanford University, PI: James F. Fries, MD, 2U01AR 052158; Boston University, PIs: Stephen M. Haley, PhD, and David Scott Tulsky, PhD, 1U01AR057929; University of California, Los Angeles, PIs: Dinesh Khanna, MD, and Brennan Spiegel, MD, MSHS, 1U01AR057936; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, 2U01AR052155; Georgetown University, Washington DC, PIs: Carol. M. Moinpour, PhD, and Arnold L. Potosky, PhD, U01AR057971; Children's Hospital Medical Center, Cincinnati, PI: Esi M. Morgan Dewitt, MD, 1U01AR057940; University of Maryland, Baltimore, PI: Lisa M. Shulman, MD, 1U01AR057967; and Duke University, PI: Kevin P. Weinfurt, PhD, 2U01AR052186). NIH Science Officers on this project have included Deborah Ader, PhD, Vanessa Ameen, MD, Susan Czajkowski, PhD, Basil Eldadah, MD, PhD, Lawrence Fine, MD, DrPH, Lawrence Fox, MD, PhD, Lynne Haverkos, MD, MPH, Thomas Hilton, PhD, Laura Lee Johnson, PhD, Michael Kozak, PhD, Peter Lyster, PhD, Donald Mattison, MD, Claudia Moy, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Ashley Wilder Smith, PhD, MPH, Susana Serrate-Sztein, MD, Ellen Werner, PhD, and James Witter, MD, PhD. This manuscript was reviewed by PROMIS reviewers before submission for external peer review. See the Web site at http://www.nihpromis.org for additional information on the PROMIS initiative.

References

1. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. Medical Care. 2007; 45(5 Suppl 1):S3–S11. [PubMed: 17443116]
2. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care. 2007; 45(5 Suppl 1):S22–S31. [PubMed: 17443115]
3. Cella D, Riley WT, Stone A, Rothrock N, Reeve B, Yount SE, et al. Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005–2008. Journal of Clinical Epidemiology. in press.

4. Bjorner JB, Kosinski M, Ware JE Jr. Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the headache impact test (HIT). Quality of Life Research. 2003; 12(8):913–933. [PubMed: 14651412]
5. Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: Item banking and computerized adaptive assessment. Quality of Life Research. 2007; 16(Suppl 1):95–108. [PubMed: 17530450]
6. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. Quality of Life Research. 2007; 16(Suppl 1): 133–141. [PubMed: 17401637]
7. Fayers PM. Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. Quality of Life Research. 2007; 16(Suppl 1):187–194. [PubMed: 17417722]
8. Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Quality of Life Research. 2010; 19(1):125–136. [PubMed: 19941077]
9. Dorans NJ. Linking scores from multiple health outcome instruments. Quality of Life Research. 2007; 16(Suppl 1):85–94. [PubMed: 17286198]
10. Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. Journal of Clinical Epidemiology. 2008; 61(10):1018–1027. [PubMed: 18455909]
11. Bjorner JB, Kosinski M, Ware JE Jr. Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. Quality of Life Research. 2003; 12(8):981–1002. [PubMed: 14651417]
12. Ware JE Jr, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlof CG, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Quality of Life Research. 2003; 12(8):935–952. [PubMed: 14651413]
13. Kitahata MM, Rodriguez B, Haubrich R, Boswell S, Mathews WC, Lederman MM, et al. Cohort profile: The Centers for AIDS Research Network of Integrated Clinical Systems. International Journal of Epidemiology. 2008; 37(5):948–955. [PubMed: 18263650]
14. Lawrence ST, Willig JH, Crane HM, Ye J, Aban I, Lober W, et al. Routine, self-administered, touch-screen, computer-based suicidal ideation assessment linked to automated response team notification in an HIV primary care setting. Clinical Infectious Diseases. 2010; 50(8):1165–1173. [PubMed: 20210646]
15. Crane HM, Lober W, Webster E, Harrington RD, Crane PK, Davis TE, et al. Routine collection of patient-reported outcomes in an HIV clinic setting: The first 100 patients. Current HIV Research. 2007; 5(1):109–118. [PubMed: 17266562]
16. 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. under review.
17. Kroenke K, Spitzer RL, Williams JB, Lowe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: A systematic review. General Hospital Psychiatry. 2010; 32(4): 345–359. [PubMed: 20633738]
18. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA. 1999; 282(18):1737–1744. [PubMed: 10568646]
19. Crane PK, Gibbons LE, Willig JH, Mugavero MJ, Lawrence ST, Schumacher JE, et al. Measuring depression and depressive symptoms in HIV-infected patients as part of routine clinical care using the 9-item Patient Health Questionnaire (PHQ-9). AIDS Care. 2010; 22(7):874–885. [PubMed: 20635252]
20. Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychology Science Quarterly. 2009; 51(2):148–180. [PubMed: 20336180]
21. Muthén, LK.; Muthén, BO. Mplus: Statistical analysis with latent variables. Los Angeles, CA: Muthén & Muthén; 1998–2007.

22. Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. Psychol Methods. 2007; 12(1):58–79. [PubMed: 17402812]
23. Forero CG, Maydeu-Olivares A. Estimation of IRT graded response models: Limited versus full information methods. Psychological Methods. 2009; 14(3):275–299. [PubMed: 19719362]
24. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No 17. 1969
25. StataCorp. Stata statistical software: Release 11. College Station, TX: StataCorp LP; 2009.
26. Muraki, E.; Bock, D. PARSCALE for Windows. Chicago: Scientific Software International; 2003.
27. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine. 2001; 16(9):606–613. [PubMed: 11556941]
28. Choi SW. Firestar: Computerized adaptive testing (CAT) simulation program for polytomous IRT models. Applied Psychological Measurement. 2009; 33(8):644–645. [PubMed: 20011609]
29. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. Journal of Rheumatology. 2009; 36(9):2061–2066. [PubMed: 19738214]
30. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). Journal of Clinical Epidemiology. 2008; 61(1):17–33. [PubMed: 18083459]
31. Crane HM, Grunfeld C, Harrington RD, Uldall KK, Ciechanowski PS, Kitahata MM. Lipoatrophy among HIV-infected patients is associated with higher levels of depression than lipohypertrophy. HIV Medicine. 2008; 9(9):780–786. [PubMed: 18754804]
32. Hansson M, Chotai J, Nordstom A, Bodlund O. Comparison of two self-rating scales to detect depression: HADS and PHQ-9. British Journal of General Practice. 2009; 59(566):e283–e288. [PubMed: 19761655]
33. Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: A systematic review. General Hospital Psychiatry. 2007; 29(5):388–395. [PubMed: 17888804]
34. Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, Revicki DA, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. Value Health. 2009; 12(4):419–429. [PubMed: 19900250]

Abbreviations

CAT	Computerized adaptive testing
CFA	Confirmatory factor analysis
CFI	Comparative Fit Index
DIF	Differential item functioning
PHQ-9	Patient Health Questionnaire from the PRIME-MD depression measure
PRO	Patient-reported outcome
PROMIS	Patient-Reported Outcome Measurement Information System
RMSEA	Root mean square error of approximation
SD	Standard deviation
SEM	Standard error of measurement
TLI	Tucker–Lewis Index
UW	University of Washington

UAB University of Alabama at Birmingham

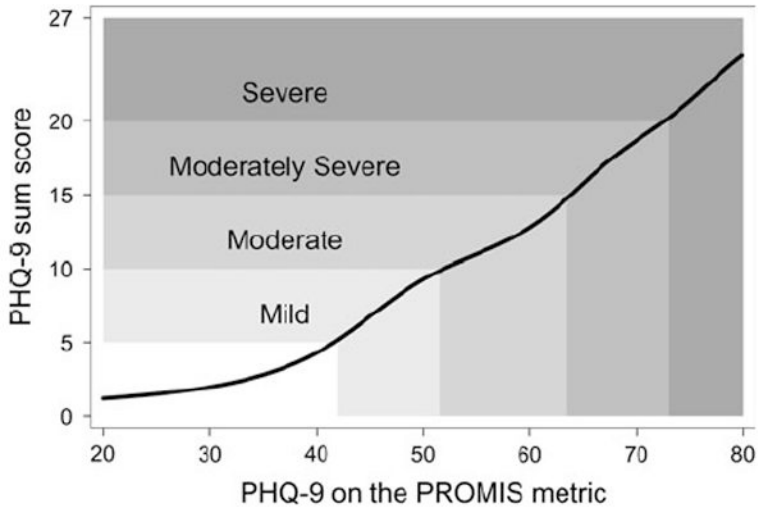

Fig. 1.

The test characteristic curve shows the most likely PHQ-9 sum score (on the *y*-axis) corresponding to an IRT-based PHQ-9 score (*x*-axis), which has been calibrated to the PROMIS metric. *Horizontal lines* indicate the published PHQ-9 cutpoints for mild (5–9), moderate (10–14), moderately severe (15–19) and severe (20–27) depression. Mild depression (PHQ-9 score of 5–9) corresponds to scores of 42–51 on the PROMIS metric, moderate depression to 52–63, moderately severe to 64–72, and severe to scores of 73 and higher

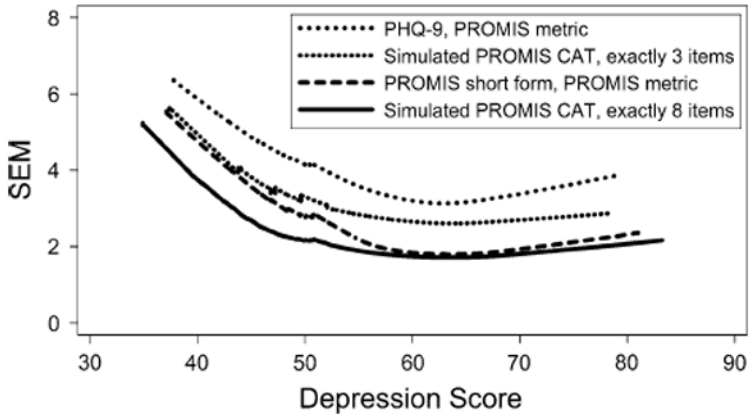

Fig. 2.

Lowess curves for the standard error of measurement (*SEM*) by depression symptom score. All the PROMIS scores, including the simulated 3-item CAT, are more precise (have smaller SEM) than the PHQ-9. In the absence of other criteria, an SEM of 0.3 SD, here equal to 3, is often used as an acceptable level of precision

Table 1

Overview of the migration from a legacy measure to PROMIS CAT administration

1	Administer both the PROMIS and legacy (here, the PHQ-9) measures to the same group of participants
2	Evaluate calibration assumptions
	a. Assess DIF
	b. Establish that they measure the same unidimensional construct
3	Using both PROMIS and legacy items, estimate the legacy parameters with the PROMIS items fixed to the PROMIS item bank parameters
4	Use the legacy items and the parameters obtained in step 3 to estimate scores. These scores will be on the PROMIS metric
5	Create a test characteristic curve to aid in interpretation of the PROMIS and legacy scores
6	Simulate CAT to determine the number of items needed for the desired level of measurement precision

Table 2

Demographic characteristics of HIV clinic participants (n = 2,178)

	University of Washington (n = 821)	**University of Alabama at Birmingham (n = 1,357)**
Age (mean (SD), range)	44 (9), 20–73	38 (10), 18–74
Male (%)	85.3	77.3
Race (%)		
White	64.1	48.4
African–American	21.1	50.0
Asian–American or Pacific Islander	3.4	0.3
American Indian	2.3	0.2
Multiracial, other or unknown	9.1	1.2

Table 3

Standard error of measurement (SEM) for depression tests calibrated to the PROMIS metric

Test	Mean (SD)	Median	Interquartile range	Range
PHQ-9, PROMIS metric	4.9 (1.4)	4.6	3.5–6.6	2.8–8.3
PROMIS short form, PROMIS metric	3.5 (1.8)	2.6	1.8–5.5	1.5–7.5
PROMIS-simulated CAT, 9 items	2.7 (1.4)	2.0	1.7–3.4	1.4–5.2
PROMIS-simulated CAT, 8 items	2.8 (1.3)	2.2	1.8–3.4	1.5–5.2
PROMIS-simulated CAT, 3 items	3.7 (1.3)	3.2	2.7–5.6	2.3–5.6

Note that the data presented in the table refer to means and standard deviations of standard errors of measurement; smaller is better. The PROMIS metric uses a SD of 10; on the unadjusted IRT logit scale, then, a mean SEM of 4.9 corresponds to a mean SEM of 0.49 standard theta units