Published in final edited form as:
J Psychiatr Res. 2014 September ; 56: 112–119. doi:10.1016/j.jpsychires.2014.05.010.

Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study

Paul A. Pilkonis, **Lan Yu**, **Nathan E. Dodds**, **Kelly L. Johnston**, **Catherine C. Maihoefer**, and **Suzanne M. Lawrence**
Department of Psychiatry, University of Pittsburgh School of Medicine

Abstract

The Patient-Reported Outcomes Measurement Information System (PROMIS®) is an NIH Roadmap initiative devoted to developing better measurement tools for assessing constructs relevant to the clinical investigation and treatment of all diseases—constructs such as pain, fatigue, emotional distress, sleep, physical functioning, and social participation. Following creation of item banks for these constructs, our priority has been to validate them, most often in short-term observational studies. We report here on a three-month prospective observational study with depressed outpatients in the early stages of a new treatment episode (with assessments at intake, one-month follow-up, and three-month follow-up). The protocol was designed to compare the psychometric properties of the PROMIS depression item bank (administered as a computerized adaptive test, CAT) with two legacy self-report instruments: the Center for Epidemiological Studies Depression scale (CESD; Radloff, 1977) and the Patient Health Questionnaire (PHQ-9; Spitzer et al., 1999). PROMIS depression demonstrated strong convergent validity with the CESD and the PHQ-9 (with correlations in a range from .72 to .84 across all time points), as well as responsiveness to change when characterizing symptom severity in a clinical outpatient sample. Identification of patients as “recovered” varied across the measures, with the PHQ-9 being the most conservative. The use of calibrations based on models from item response theory (IRT) provides advantages for PROMIS depression both psychometrically (creating the possibility of adaptive testing, providing a broader effective range of measurement, and generating greater precision) and practically (these psychometric advantages can be achieved with fewer items—a median of 4 items administered by CAT—resulting in less patient burden).

Keywords

depression; item response theory; measurement; self-report; patient-reported outcomes

1. Introduction

The Patient-Reported Outcomes Measurement Information System (PROMIS®) is an NIH Roadmap initiative devoted to developing better measurement tools for assessing constructs

Correspondence concerning this article should be addressed to Paul A. Pilkonis, Western Psychiatric Institute and Clinic, 3811 O’Hara Street, Pittsburgh, PA 15213. pilkonispa@upmc.edu. Phone: 412-246-5833. Fax: 412-246-5840.

relevant to the clinical investigation and treatment of all diseases—constructs such as pain, fatigue, emotional distress, sleep, physical functioning, and social participation (Buysse et al., 2010; Cella et al., 2010; Cella et al., 2007b; Fries et al., 2009; Fries et al., 2014; Pilkonis et al., 2011; Revicki et al., 2009). PROMIS has created and refined a comprehensive methodology for developing item banks of these health-related constructs using both qualitative and quantitative techniques and modern psychometric methods (item response theory, IRT) (Cella et al., 2007a; Cella et al., 2010; Hilton, 2011; Reeve et al., 2007). These item banks encompass physical, mental, and social health, consistent with the World Health Organization's tripartite framework (Cella et al., 2007a; World Health Organization, 2007).

The use of models from IRT to calibrate items not only results in greater precision at the item and test levels but also promotes greater flexibility in test administration. For example, items can be administered as computerized adaptive tests (CATs), or static short forms can be created and tailored for samples with different levels of severity of the construct being assessed. Analyses of potential differential item functioning due to gender, age, and educational attainment were performed during the development of the item banks to ensure that items performed comparably regardless of variations in these background characteristics. In general, experience with CAT suggests that the PROMIS depression item bank provides excellent precision with 4–6 items (Choi et al., 2010). A generic 8-item short form is also available, and this short form was one of the cross-cutting dimensional measures used in the DSM-5 field trials, where its feasibility was established and where it performed well with regard to test-retest reliability (Narrow et al., 2013). Following creation of the item banks, our priority has been to validate them, most often in short-term observational studies. These studies allow us to examine the psychometric properties of the item banks, their responsiveness to change, their relationships to clinically significant benchmarks of improvement, and their similarities and differences when compared with other commonly used instruments.

We report here on a prospective observational study with depressed outpatients in the early stages of a new treatment episode. For this purpose, all participants completed study assessments at three points: baseline (T1, as close to the beginning of treatment as possible but no later than four months after its start), one month following baseline (T2), and three months following baseline (T3). The protocol was designed to compare the psychometric properties of the PROMIS depression item bank (administered as a CAT) with two legacy self-report instruments: the Center for Epidemiological Studies Depression scale (CESD; Radloff, 1977) and the Patient Health Questionnaire (PHQ-9; Spitzer et al., 1999).

The study was not intended to evaluate treatment effectiveness. Rather, the main consideration was to conduct a study involving established treatments that would allow us to investigate the operating characteristics of the different measures of depression over a time frame (three months) consistent with the design of clinical trials and comparative effectiveness research. Regardless of their impact in the aggregate, treatments for depression generate considerable variability in individual outcomes, and this variability was desirable for examining psychometric issues. In our setting, the most common form of outpatient treatment for depression is a combination of antidepressant medication and supportive psychotherapy (both individual and group therapies), with smaller proportions of patients receiving medication only or psychotherapy only. No untreated or control group was included.

There have been other attempts to link PROMIS depression to legacy measures for depression. The PROsetta Stone project (Choi et al., 2012) was designed specifically to create "cross-walks" between PROMIS measures in several domains and commonly used measures (most often developed using classical test theory) in those same domains. The PROsetta Stone web site (see http://www.prosettastone.org/Pages/default.aspx) provides a conversion table (Appendix Table 7) from raw CESD scores to PROMIS depression scores. The PROMIS depression equivalent for the CESD threshold of 16 is 56.2; for the CESD threshold of 21, it is 59.1. (Note that PROMIS depression is scored with a T-score metric in which the mean of the general population is 50, with a standard deviation of 10.)

The PROsetta Stone web site also provides a conversion table (Appendix Table 46) from raw PHQ-9 scores to PROMIS depression scores. The PROMIS depression equivalent for the PHQ-9 threshold of 5 (mild depression) is 52.5; for the threshold of 10 (moderate depression), 59.9; for the threshold of 15 (moderately severe depression), 65.8; and for the threshold of 20 (severe depression), 71.5. Gibbons et al. (2011) also reported analyses linking PROMIS depression and the PHQ-9 in a sample of HIV patients. Their results were generally comparable to the PROsetta Stone linkages. However, there was some discrepancy at the mild end of the PHQ-9 where they found rather low PROMIS depression scores to be equivalent: "Mild depression (PHQ-9 score of 5–9) corresponds to scores of 42–51 on the PROMIS metric, moderate depression [10–14] to 52–63, moderately severe [15–19] to 64–72, and severe [20+] to scores of 73 and higher" (figure caption, p. 1353). In general, thresholds suggesting depression of some clinical significance (CESD = 21, PHQ-9 = 10) have been linked to a PROMIS score of about 60, the usual threshold used clinically with the T-score metric (1 SD above the mean).

Finally, in a study using two different IRT linking methods, Olino et al. (2013) compared the Beck Depression Inventory (Beck et al., 1961) the CESD, and the PROMIS depression item banks in a community sample of adolescents. Among the three measures, PROMIS depression provided information over the widest range of symptom severity while demonstrating the highest level of precision. This result was especially true for the full PROMIS depression item bank of 28 items, but it also applied to the PROMIS depression short form of 8 items, which is considerably briefer than either the BDI or the CESD.

2. Method

2.1. Inclusion criteria

Men and women 18 years and older who were able to read and understand English and able and willing to give informed consent were enrolled in the protocol. They were required to be within the first four months of outpatient treatment for major depressive disorder (MDD) at Western Psychiatric Institute and Clinic (WPIC) and its affiliates. To ensure that participants were not too close to the floor for depression when beginning the protocol (and thus unable to show further change), we required a minimum score of 12 on the 17-item Hamilton Rating Scale for Depression (Hamilton, 1960).

2.2. Exclusion criteria

Clinical exclusions were dementia, other cognitive impairment, or a history of any psychotic or bipolar disorder that might compromise the validity of self-reports or interfere with questionnaire completion. In addition, we excluded patients with organic affective syndromes (i.e., mood disorder secondary to a general medical condition or substance-induced mood disorder) and with major medical conditions that may have a significant impact on the central nervous system (e.g., Parkinson's disease, stroke, multiple sclerosis, systemic lupus erythematosus). We also excluded patients with a history of continuous care for one year or more in the mental health care system within the past five years to eliminate patients with chronic presentations that were less likely to demonstrate change in the short term.

2.3. Sample

The sample included 194 patients with a mean age of 48 (*SD* = 19), ranging from 18 to 83. At baseline, the median number of weeks in treatment was 4.9, with 44% of patients having started treatment within the past month. Sixty-four percent of patients received a combination of medication and psychotherapy, 28% received medication only, and 8% received psychotherapy only. In terms of gender, ethnicity, and race, 74% of the sample was female; 6% was Hispanic; and 80% was White, 14% was African American, 2% was American Indian, and 2% was Asian American (with 2% unknown). In terms of educational level, 27% of the sample had a high school education or less, 55% had some college education or a college degree, and 17% had an advanced degree. In terms of household income, 31% of the sample had income less than $20,000, 36% had income between $20,000 and $49,999, 21% had income between $50,000 and $99,999, and 8% had income of $100,000 or more (with 5% unknown).

2.4. Test administration

PROMIS measures were administered as CATs. The CAT algorithm required a minimum of 4 items and a maximum of 12, and it stopped the test when the standard error was less than .30. Across all tests at T1, T2, and T3, participants received an average of 4.8 items for the depression CAT (SD = 1.9), with a median of 4 items. Legacy measures were also administered by computer but in their usual static form, i.e., all items were presented to participants—20 items for the CESD and 9 items for the PHQ-9. In all formats, items were displayed one at a time using Assessment Center, the PROMIS electronic testing platform (see http://www.assessmentcenter.net).

2.5. Statistical methods

The measurement issues of primary interest were convergent validity and responsiveness to change. To investigate convergent validity, we performed two analyses. First, we examined Pearson correlations between PROMIS depression and the legacy measures (CESD and PHQ-9), both cross-sectionally and longitudinally. Second, we examined the PROMIS equivalents of conventional thresholds and ranges of depression (e.g., mild, moderate, and severe depression) on the legacy measures at all time points.

To investigate responsiveness to change, we performed three analyses. First, we examined effect sizes across the PROMIS and legacy measures at the two follow-up evaluations, one month (T2) and three months (T3) after the initial assessment. Second, we examined clinically meaningful differences across the measures as anchored by patient ratings of global improvement (from "very much improved" to "worse") at the final follow-up assessment (T3). For this purpose, we report both mean changes and effect sizes associated with each level of improvement, with a goal of identifying the magnitude of changes on each measure that are seen as important by patients. Third, we examined categorical outcomes for "recovery" using two different methods. The first method (Jacobson and Truax, 1991) used (a) the reliable change index (RCI), which identifies the amount of change necessary to be confident that such change exceeds measurement error, and (b) computation of a cutoff score distinguishing normative and clinical samples. After these two quantities are computed, patients can be identified as "recovered" (those whose improvement exceeds the RCI and whose final scores place them below the cutoff between normative and clinical samples), "improved" (change exceeding the RCI but final status above the cutoff), "unchanged" (change less than the RCI in either a positive or negative direction), and "deteriorated" (worsening exceeding the RCI).

The second method for examining recovery used absolute thresholds to define recovery. For PROMIS depression, this was a final score less than 55 on the PROMIS T-score metric, i.e., final status less than ½ SD above the normative mean (with an initial score above the usual clinical threshold of 60 or +1 SD above the mean for T-scores). For the CESD, recovery required a final score less than 16 (with an initial score above 20). For the PHQ-9, recovery required a final score less than 5 (with an initial score above 9). We also examined "subthreshold" cases in this same way, i.e., patients with initial PROMIS depression scores between 55 and 60, initial CESD scores of 16–20, and initial PHQ-9 scores of 5–9, to see what proportions of them fell below conventional thresholds for recovery at the final assessment (T3). With both methods, we used McNemar's test for differences in proportions to examine the statistical significance of variations in the proportions of recovery.

After the RCI outcome categories had been identified, we examined agreement on these categories in pairwise comparisons between the measures using weighted kappa coefficients (given the ordinal nature of the categories from "recovered" to "deteriorated"). Agreement between measures is informative, but we also investigated the validity of the different outcomes across the different instruments by examining relationships between "recovery" and other clinically significant indicators. These indicators (collected at the final T3 assessment) included Global Assessment of Functioning (GAF) scores and outcomes on other PROMIS item banks relevant to depression, e.g., anxiety, sleep, fatigue, and social functioning (both in social roles and in discretionary social activities).

2.6. Human use issues

The study was carried out in accordance with the latest version of the Declaration of Helsinki and was reviewed by the University of Pittsburgh Institutional Review Board. Informed consent was obtained from participants after all procedures had been fully explained.

3. Results

3.1. Descriptive statistics

Cronbach's alpha was used to compute the reliabilities of the legacy measures at baseline, which were .86 for the CESD and .81 for the PHQ-9. For measures derived from IRT models, test information (and its converse, standard error, SE) varies along the spectrum of severity of the construct being assessed. The reliability of PROMIS depression was .92 when calculated as

$$\text{Reliability}=1-\frac{SE^2_{baseline}}{SD^2_{baseline}}$$

where $\text{SE}_{\text{baseline}}$ is the median of the SE of PROMIS depression in a range from −3 to +3 standard deviations.

Table 1 displays the mean scores and standard deviations across the three time points for the PROMIS and legacy measures. In the aggregate, one sees the expected trend on all measures —decreases in mean scores over time, all of which were moderate to large when transformed into effect sizes (see section 3.3.1 below). Inspection of the frequency distributions for all measures at all time points showed that PROMIS depression scores most closely approximated a normal distribution, consistent with the goal of capturing a broad effective range of measurement throughout the full spectrum of severity. Average skewness across the three time points was .11 for PROMIS depression, .40 for the CESD, and .60 for the PHQ-9, with increasing skewness over time for the legacy measures as patients became less depressed and scores piled up at the floor. PROMIS depression demonstrated less compression at the floor, even at lower levels of depression for the sample as a whole. Also, average kurtosis across the three time points was −.05 for PROMIS depression, −.35 for the CESD, and −.27 for the PHQ-9. Thus, the legacy measures had more positively skewed distributions (as a function of more pronounced floor effects) and flatter distributions (negative kurtosis).

3.2. Convergent validity

The correlations between PROMIS depression and the legacy measures were significant at all three time points, and the magnitudes were large, in a range from .72 to .84 (see Table 2). The legacy measures for depression, the CESD and the PHQ-9, have a history of psychometric research that has led to standard interpretations of ranges of scores associated with different levels of severity. For the CESD, we examined the nonclinical range of 0–15, a subthreshold range of 16–20, and a clinical range of 21 and higher. For the PHQ-9, we examined the asymptomatic range of 0–4, the mild symptom range of 5–9, the moderate symptom range of 10–14, the moderately severe symptom range of 15–19, and the severe symptom range of 20 and higher. Table 3 summarizes the PROMIS equivalents of these conventional ranges used with the CESD and the PHQ-9 for classifying different levels of depression.

The nonclinical range of the CESD produced PROMIS mean scores that were normative, the subthreshold range produced mean scores that were elevated approximately one-half standard deviation above the mean, and the clinical range produced mean scores that were above the usual threshold of concern for T-scores, i.e., a score of 60, which is one standard deviation above the mean. Also, the PROMIS scores linked to the same tiers of the CESD appeared to be stable across the three time points, even as the sample became less symptomatic in general. Thus, on the CESD, patients scoring in the nonclinical range of 0–15 across the three assessment points produced average PROMIS scores of 51.2, 48.0, and 48.3. For the subthreshold range of 16–20, the average PROMIS scores were 53.2, 55.2, and 53.9 across the three assessment points. For the clinical range of 21 or higher, the average PROMIS scores were 64.5, 63.2, and 64.0 across the three assessment points.

The results for the mapping between PHQ-9 ranges and PROMIS depression mean scores produced values that were also consistent with the intent of the PHQ-9 and that remained consistent over time, results comparable to those seen with the CESD. On the PHQ-9, patients scoring in the asymptomatic range of 0–4 across the three assessment points produced average PROMIS scores of 47.7, 48.1, and 47.6, scores that were slightly below the normative T-score of 50. For the mild symptom range of 5–9, the average PROMIS scores were 55.9, 55.2, and 54.7 across the three assessment points, about one-half standard deviation above the mean. For the moderate symptom range of 10–14, the average PROMIS scores were 60.7, 59.2, and 60.2, consistent with the usual threshold of concern for T-scores one standard deviation above the mean. For the moderately severe symptom range of 15–19, the average PROMIS scores were 66.4, 65.8, and 65.5 across the three assessment points, representing another increment of approximately one-half standard deviation. For the severe symptom range of 20 and higher, the average PROMIS scores were 70.7, 71.1, and 74.9 across the three assessment points.

3.3. Responsiveness to Change

3.3.1. Group-level analyses—All the depression measures produced similar effect sizes at the 1-month follow-up assessment (T2): PHQ-9, 0.69; CESD, 0.63; and PROMIS depression, 0.56. There was greater separation, however, at the 3-month follow-up (T3) when the rank order of the effect sizes was CESD, 1.06; PHQ-9, 0.98; and PROMIS depression, 0.84. To examine further the comparability of effect sizes at the 3-month follow-up assessment, normalized change scores (*z*) were computed for each participant for PROMIS depression, CESD, and PHQ-9. Paired t-tests were performed between the measures using these *z*-scores. The difference between the CESD and PHQ-9 was not statistically significant, but the differences between PROMIS depression and the CESD ($p < .01$) and PROMIS depression and the PHQ-9 ($p < .02$) were significant, with greater (normalized) change on the CESD and the PHQ-9, consistent with the differences in the group-based effect sizes.

Table 4 summarizes the outcomes on the three measures of depression when they were linked to patient ratings of global improvement at the final assessment. With all three measures of depression, global ratings of “very much” improved were associated with effect sizes of 1.5–1.7, and global ratings of “much” improved were associated with effect sizes of

1.0–1.4. In general, the effect sizes produced by the three measures of depression were similar, with a tendency for the effect sizes generated by the CESD to be the largest.

3.3.2. Person-level analyses—We investigated responsiveness to change at the individual level in two ways. First, we used the reliable change index (RCI) method proposed by Jacobson and Truax (1991). To calculate cutoff scores distinguishing between clinical and normative samples, data from both kinds of samples are required. Means and SDs on all the measures from the current sample at intake were used as estimates for clinical samples. During their development, PROMIS measures were centered (using the T-score metric) on data from the general population; thus, the normative mean for PROMIS depression is 50, with an SD of 10. Crawford et al. (2011) were the source for normative data on the CESD (*M* = 10.2, SD = 9.7), whereas Kocalevent et al. (2013) reported normative data for the PHQ-9 (*M* = 2.9, SD = 3.5). Based on these estimates, the cutoff score for PROMIS depression was 56.6; for CESD, 19.5; and for PHQ-9, 7.0. Final status below these cutoffs and a change score exceeding the RCI were required to be classified as "recovered." The raw change score exceeding the RCI for PROMIS depression was 6.4; for the CESD, 10.6; and for the PHQ-9, 6.7.

Table 5 summarizes classifications on all three measures at the 3-month follow-up assessment (T3) using the mutually exclusive categories of "recovered," "improved," "unchanged," and "deteriorated." For depression, the CESD provided the most liberal estimate of "recovered" (42%) and the PHQ-9 produced the most stringent (29%), with PROMIS depression (39%) closer to the CESD than the PHQ-9. The paired difference between PROMIS Depression and the CESD was not statistically significant, but the proportion of recovered patients on the PHQ-9 was significantly lower than that from both PROMIS depression ($p < .04$) and the CESD ($p < .01$). Thus, the different instruments provided different "snapshots" of these clinically relevant categories, with the CESD and PROMIS depression being more generous than the PHQ-9.

Because patients could be classified differently on the different measures, we calculated weighted kappa coefficients to examine concordance in the four ordinal classifications between measures. The kappas were very similar in the three possible pairwise comparisons: CESD versus PROMIS depression (k = .61, percent agreement = 70%), CESD versus PHQ-9 (k = .58, percent agreement = 69%), and PROMIS depression versus PHQ-9 (k = . 58, percent agreement = 67%). These pairwise comparisons suggested that about 70% of cases will be classified the same by any two of the instruments when using the four categories of outcome.

The second method used for person-level analyses of change relied on a priori thresholds to define "recovery." For PROMIS depression, this threshold was a final score less than 55 on the PROMIS T-score metric, i.e., less than one-half SD above the normative mean. Note that this threshold was more stringent than the cutoff score (56.6) used in the RCI analyses described above. For the CESD, recovery required a final score less than 16, and for the PHQ-9, recovery required a final score less than 5. Both of these thresholds are consistent with conventional uses of these measures, and they are also more stringent than the cutoff scores used in the RCI analyses.

We examined outcomes in two ways with these rationally defined thresholds. First, we identified participants with clinically significant elevations on the depression measures. For PROMIS depression, this was an initial score above the usual clinical threshold of 60 or +1 SD above the mean for T-scores. For the CESD, this was an initial score of 21 or higher, and for the PHQ-9, this was an initial score of 11 or higher. For this subsample, we computed the percentages of patients who fell below the absolute thresholds for recovery at the final follow-up assessment (T3, see Table 6). Second, we repeated this analysis with the "subthreshold" cases, i.e., patients with initial PROMIS depression scores between 55 and 60, initial CESD scores of 16–20, and initial PHQ-9 scores of 5–9 to see what proportions of them fell below absolute thresholds for recovery at the final assessment.

As Table 6 illustrates (and consistent with the RCI analysis results in Table 5), there were differences in recovery rates depending on the instrument examined. In the subsample of patients with clinically significant elevations at the baseline assessment, the percentages for recovery varied from 41% on PROMIS depression to 32% on the PHQ-9, with the CESD at 39%. The result with PROMIS depression was not statistically different from that with the CESD, but again, the proportion of recovered patients on the PHQ-9 was significantly lower than that from both PROMIS depression ($p < .04$) and the CESD ($p < .03$). Using this approach, the PHQ-9 was again the most conservative of the measures. For the subsample of subthreshold cases, the rates of recovery were higher (given that these patients were closer to the final score required for recovery when first assessed), but again, the percentages varied across the measures: 79% for the CESD, 67% for PROMIS depression, and 51% for the PHQ-9. These differences, however, were not significantly different because of the small sample sizes available for these comparisons.

Investigating the performance of the different measures in identifying patients who were recovered is informative, but we also wanted to link such performance to other measures of outcome in a further effort at validation. For this purpose, we performed one-way analyses of variance using the four categories of outcome (recovered, improved, unchanged, and deteriorated) from the RCI analyses as the grouping (independent) variable for PROMIS depression, the CESD, and the PHQ-9. As dependent variables at the 3-month follow-up assessment (T3), we used other measures that are clinically relevant to depression: the Global Assessment of Functioning (GAF) scale and other PROMIS CATs administered as part of the current protocol. These CATs assessed anxiety, sleep disturbance, sleep-related (daytime) impairment, fatigue, and social functioning (both in social roles and in discretionary social activities).

Table 7 summarizes the results for the GAF and the PROMIS fatigue CAT, which was representative of the findings from all the CATs. With both of these outcome measures, the most important distinction occurred between patients classified as "recovered" and all others. With the GAF, the results indicated that the final rating (T3) was about 70 for patients who recovered from their depression (on any of the three measures); 70 is the best score in the decile (61–70) for "mild" symptoms, indicating that the patient was "generally functioning pretty well." For patients in the other three categories, GAF scores ranged from 58 to 63, spanning the better end of the decile (51–60) for "moderate" symptoms and the poorer end of the decile (61–70) for "mild" symptoms.

With fatigue (and the other PROMIS CATs), the results indicated that scores were normative (T-score about 50) for patients who recovered from their depression (on any of the three measures). For patients in the other three categories, fatigue scores remained elevated with T-scores from 58 to 66. Post hoc comparisons of the means from the four-group, one-way ANOVAs using the Tukey-b test documented that the differences between the "recovered" group and the other three groups were statistically significant ($p < .05$) on the GAF and the PROMIS CATs, with the latter three groups (improved, unchanged, and deteriorated) generally not differing significantly among themselves (see Table 7). Again, this pattern was similar across the three measures of depression.

4. Discussion

We report here on a prospective observational study with depressed outpatients in the early stages of a new treatment episode which was designed to compare the psychometric properties of the PROMIS depression item bank (administered as a CAT) with two legacy self-report instruments: the CESD and the PHQ-9. The study allowed us to examine the psychometric properties of the measures (frequency distributions, reliabilities), their convergent validity (correlations, linkages to commonly used thresholds for severity of depression on the legacy measures), their relationships to patient ratings of global improvement in order to identify the magnitude of change associated with patient perception of important gains, and their responsiveness to change (effect sizes, proportions of patients identified as recovered by each measure).

The results demonstrated that PROMIS depression had greater reliability and scores more closely approximating a normal distribution than the two legacy measures. At the same time, PROMIS depression demonstrated strong convergent validity with the CESD and the PHQ-9. The correlations between PROMIS depression and the legacy measures were large at all three time points, in a range from .72 to .84. Also, PROMIS depression mapped, in sensible ways, onto the conventional ranges and benchmarks used with the legacy measures. The nonclinical range of the CESD produced PROMIS mean scores that were normative, the subthreshold range produced mean scores that were elevated approximately one-half standard deviation above the mean, and the clinical range produced mean scores that were above the usual threshold of concern, i.e., a T-score of 60. Also, the average PROMIS scores linked to the same tiers of the CESD were quite similar across the three time points. These findings support both the validity of the tiers represented in the CESD and the ability of the PROMIS CAT for depression to detect the same "signal" in a consistent way. The results for the mapping between PHQ-9 ranges and mean PROMIS depression scores produced values that were also consistent with the clinical intent of the PHQ-9 and that remained similar over time.

One unexpected finding emerged at the three-month follow-up assessment when PROMIS depression displayed the smallest effect size of the three measures. Given the apparent psychometric advantages of PROMIS depression (greater reliability; a more normal distribution of scores, presumably allowing for greater sensitivity to change), one might expect that larger effect sizes would appear as a result of larger pre-post mean differences—the numerator in the effect size calculation. Inspection of the frequency distributions suggests, however, that floor effects with the legacy measures decreased their variance, even at intake, and that this decreased variance (the denominator in the effect size calculations) contributed to the larger effect sizes of the legacies. Such a result raises the sobering possibility that commonly used measures of depression may overestimate effect sizes in some circumstances because of the presence of floor effects and their positive skewness.

There was variability across the measures when they were used to identify categorical outcomes (“recovered”) at the individual level. The PHQ-9 was the most conservative in this regard, which may reflect, in part, the different content of the measures. The PHQ-9 is linked specifically to the DSM criteria for major depressive disorder, which are relatively “difficult” items to endorse and, once present, may create greater challenges for full recovery. The CESD and PROMIS depression focus more generally on affective and cognitive aspects of poor mood, and these facets of depression may be more amenable to change.

As a general psychometric concern, the issue of lessening patient burden and increasing the efficiency of assessment deserves discussion. Across all test administrations in the present study, participants received an average of 4.8 items (SD = 1.9) for the PROMIS depression CAT, with a median of 4 items, convincing evidence of the efficiency of the CAT. The PHQ-9 consists, of course, of 9 items, and some variants have utilized just the first 2 items for screening purposes. Thus, both PROMIS depression and the PHQ-9 impose less burden that the CESD, which is most commonly used in its 20-item version.

It is also necessary to acknowledge the limitations of the current work. The most important are limitations on generalizability. Given the sampling frame, the current study reflects the pros and cons of the measures in an outpatient clinical sample in an academic health center providing secondary and tertiary care. The mandate for the PROMIS measures is to provide a common metric for assessing symptoms, functioning, and health-related quality of life across chronic diseases (both medical and psychiatric) and along the full spectrum of severity (characteristic of community as well as clinical samples). The present investigation reflects just one segment of this ambitious agenda. Given that validation is an evolving process (rather than a single outcome), it remains important to continue to test the operating characteristics of PROMIS depression and other legacy instruments in a broad range of contexts—psychiatric, medical, and epidemiological.

In summary, the results presented here suggest that PROMIS depression can be a valuable addition to the family of measures used for the assessment of depression. Using an item bank calibrated with models from IRT provides advantages both psychometrically (creating the possibility of adaptive testing, providing a broader effective range of measurement, and generating greater precision) and practically (these psychometric advantages can be achieved with fewer items and less patient burden). Therefore, we encourage clinical researchers to consider adoption of the PROMIS depression item bank, especially when used as a CAT.

References

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]

Buysse DJ, Yu L, Moul DE, Germain A, Stover A, Dodds NE, Johnston KL, Shablesky-Cade MA, Pilkonis PA. Development and validation of patient-reported outcome measure for sleep disturbance and sleep-related impairments. Sleep. 2010; 33:781–792. [PubMed: 20550019]

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. 2007a; 16:133–144. [PubMed: 17401637]

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, Lai JS, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R. The Patient-Reported Outcomes Measurement Information System (PROMIS®) developed and tested its first wave of adult self-reported outcome item banks: 2005–2008. Journal of Clinical Epidemiology. 2010; 63:1179–1194. [PubMed: 20685078]

Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS®): progress of an NIH Roadmap cooperative group during its first two years. Medical Care. 2007b; 45:S3–S11. [PubMed: 17443116]

Choi, SW.; Podrabsky, T.; McKinney, N.; Schalet, BD.; Cook, K.; Cella, D. PROSetta® Stone Analysis Report: a Rosetta Stone for Patient Reported Outcomes (Vol. 1). Chicago, IL: Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University; 2012.

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. Qual Life Res. 2010; 19:125–136. [PubMed: 19941077]

Crawford J, Cayley C, Lovibond PF, Wilson PH, Hartley C. Percentile norms and accompanying interval estimates from an Australian general adult population sample for self-report mood scales (BAI, BDI, CRSD, CES-D, DASS, DASS-21, STAI-X, STAI-Y, SRDS, and SRAS). Australian Psychologist. 2011; 46:3–14.

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; 36:2061–2066. [PubMed: 19738214]

Fries JF, Witter J, Rose M, Cella D, Khanna D, Morgan-Dewitt E. Item response theory, computerized adaptive testing, and PROMIS: Assessment of Physical Function. Journal of Rheumatology. 2014; 41:153–158. [PubMed: 24241485]

Gibbons LE, Feldman BJ, Crane HM, Mugavero M, Willig JH, Patrick D, Schuumacher 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; 20:1349–1357. [PubMed: 21409516]

Hamilton M. A rating scale for depression. Journal of Neurology, Neurosurgery, and Psychiatry. 1960:23.

Hilton TF. The promise of PROMIS® for addiction. Drug and Alcohol Dependence. 2011; 119:229–234. [PubMed: 22238781]

Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Clinical Psychology. 1991; 59:12–19.

Kocalevent RD, Hinz A, Brahler E. Standardization of the depression screener patient health questionnaire (PHQ-9) in the general population. General Hospital Psychiatry. 2013; 35:551–555. [PubMed: 23664569]

Narrow WE, Clarke DE, Kuramoto SJ, Kraemer HC, Kupfer DJ, Greiner L, Regier DA. DSM-5 field trials in the United States and Canada, Part III: Development and reliability of a cross-cutting symptom assessment for DSM-5. American Journal of Psychiatry. 2013; 170:71–82. [PubMed: 23111499]

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; 1:1267–1277. [PubMed: 23686132]

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; 18:263–283. [PubMed: 21697139]

Radloff LS. A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977; 1:385–401.

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Cella D, group obotPc. Psychometric evaluation and calibration of Health-Related Quality of Life item banks: Plans for the Patient-Reported Outcome Measurement Information System (PROMIS®). Medical Care. 2007; 45:S22–S31. [PubMed: 17443115]

Revicki D, Chen W, Harnam N, Cook K, Amtmann D, Callahan LF, Jensen MP, Keefe FJ. Development and psychometric analysis of the PROMIS® pain behavior item bank. Pain. 2009; 146:158–169. [PubMed: 19683873]

Spitzer R, Kroenke K, Williams J. Validation and utility of a self-report version of PRIME-MD: The PHQ Primary Care Study. Journal of the American Medical Association. 1999; 282:1737–1744. [PubMed: 10568646]

World Health Organization. Constitution of the World Health Organization: Basic documents (46th ed.). Geneva: Author; 2007.

Table 1

Mean Scores and Standard Deviations for PROMIS Depression and Legacy Measures

	Baseline (T1)			1-Month (T2)			3-Month (T3)		
Measure	**M**	**SD**	**n**	**M**	**SD**	**n**	**M**	**SD**	**n**
PROMIS Depression	62.0	8.1	194	57.5	9.1	186	55.2	9.9	187
CESD	29.3	10.2	191	22.7	11.4	183	18.4	11.7	185
PHQ-9	13.3	5.5	193	9.5	5.9	186	7.8	6.1	186

Table 2

Correlations between PROMIS Depression, CESD, and PHQ-9 within and across time points

	PROMIS Depression			CESD			PHQ-9		
Measure	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
PROMIS Depression									
Baseline (T1)	---								
1-Month (T2)	.58	---							
3-Month (T3)	.53	.68	---						
CESD									
Baseline (T1)	**.82**	.56	.46	---					
1-Month (T2)	.52	**.84**	.60	.58	---				
3-Month (T3)	.42	.58	**.84**	.44	.64	---			
PHQ-9									
Baseline (T1)	**.72**	.41	.33	**.81**	.45	.32	---		
1-Month (T2)	.53	**.76**	.60	.58	**.84**	.64	.55	---	
3-Month (T3)	.40	.60	**.81**	.41	.61	**.89**	.38	.72	---

Note. All correlations are statistically significant. Convergent correlations between the measures at the same time points are bolded.

Table 3

PROMIS Depression Equivalents for Clinical Tiers of the CESD and PHQ-9

	Baseline (T1)					1-Month (T2)					3-Month (T3)				
		CESD		PROMIS Depression			CESD		PROMIS Depression			CESD		PROMIS Depression	
CESD score	**n (%)**	**M**	**SD**	**M**	**SD**	**n (%)**	**M**	**SD**	**M**	**SD**	**n (%)**	**M**	**SD**	**M**	**SD**
0–15	18 (9)	11.7	2.4	51.2	4.0	53 (29)	9.4	3.4	48.0	6.3	85 (46)	8.4	3.7	48.3	6.7
16–20	21 (11)	18.1	1.5	53.2	3.9	31 (17)	18.6	1.4	55.2	6.0	32 (17)	17.6	1.3	53.9	4.9
21+	152 (80)	32.9	8.0	64.5	7.0	99 (54)	31.1	7.8	63.2	6.4	68 (37)	31.3	8.0	64.0	7.6
		PHQ-9		PROMIS Depression			PHQ-9		PROMIS Depression			PHQ-9		PROMIS Depression	
PHQ-9 score	n (%)	M	SD	M	SD	n (%)	M	SD	M	SD	n (%)	M	SD	M	SD
0–4	5 (3)	2.8	0.8	47.7	4.8	37 (20)	2.6	1.4	48.1	7.9	72 (39)	2.2	1.5	47.6	6.7
5–9	52 (27)	7.3	1.2	55.9	5.1	68 (37)	6.7	1.3	55.2	6.0	56 (30)	7.0	1.5	54.7	5.7
10–14	57 (30)	12.1	1.4	60.7	6.1	42 (23)	11.5	1.0	59.2	5.2	26 (14)	12.2	1.5	60.2	5.5
15–19	49 (25)	16.8	1.4	66.4	5.8	23 (12)	16.4	1.3	65.8	5.9	22 (12)	16.4	1.1	65.5	5.4
20+	30 (16)	22.1	2.2	70.7	6.4	16 (9)	22.0	1.9	71.1	6.3	10 (5)	22.4	2.1	74.9	5.0

Table 4

Difference Scores and Effect Sizes Associated with Patient Ratings of Clinical Improvement at 3 Months (T3)

	PROMIS Depression			CESD			PHQ-9		
Global Improvement	**n**	**Mean Change**	**Effect Size**	**n**	**Mean Change**	**Effect Size**	**n**	**Mean Change**	**Effect Size**
Very much improved	69	−12.0	1.48	69	−17.6	1.72	69	−9.5	1.67
Much improved	50	−7.7	1.04	48	−12.6	1.35	50	−5.6	1.13
Minimally improved	41	−2.2	0.29	41	−4.6	0.45	41	−1.6	0.29
Same	13	−1.3	0.15	13	−1.2	0.11	12	−1.3	0.19
Worse	12	1.1	−0.13	12	2.3	−0.22	12	0.6	−0.12

Table 5

Categorical Outcomes Using the Reliable Change Index (RCI) Method

	PROMIS Depression		CESD		PHQ-9	
Outcome	**n**	**%**	**n**	**%**	**n**	**%**
Recovered	73	39%	77	42%	54	29%
Improved	16	9%	11	6%	20	11%
Unchanged	90	48%	89	49%	107	58%
Deteriorated	8	4%	5	3%	5	3%

Table 6

Outcomes for Patients Using Absolute Thresholds for Recovery at 3 Months (T3)

		PROMIS Depression		CESD		PHQ-9	
Status at Baseline (T1)	**Status at 3-Month (T3)**	**n**	**%**	**n**	**%**	**n**	**%**
Depressed	< Threshold	45	41%	57	39%	42	32%
	≥ Threshold	64	59%	88	61%	88	68%
Subthreshold	< Threshold	24	67%	15	79%	26	51%
	≥ Threshold	12	33%	4	21%	25	49%

Note. For PROMIS depression, the threshold for recovery was a final score less than 55 on the PROMIS T-Score metric, i.e., less than one-half standard deviation above the normative mean. For the CESD, the threshold for recovery was a final score less than 16, and for the PHQ-9, a final score less than 5.

Table 7

Mean Global Assessment of Functioning (GAF) and PROMIS Fatigue Scores at 3 Months (T3)

Outcome	**PROMIS Depression**	**CESD**	**PHQ-9**
		GAF	
Recovered	69.3_{a}	69.5_{a}	70.6_{a}
Improved	59.4_{b}	62.7_{ab}	63.1_{b}
Unchanged	62.3_{b}	61.1_{b}	62.1_{b}
Deteriorated	57.6_{b}	59.4_{b}	61.2_{b}
		PROMIS Fatigue	
Recovered	50.9_{a}	50.2_{a}	48.6_{a}
Improved	62.4_{b}	66.2_{b}	60.8_{b}
Unchanged	58.1_{b}	59.0_{b}	58.2_{b}
Deteriorated	61.6_{b}	61.6_{b}	61.1_{b}

Note. Outcomes regarding recovery are those identified by the Reliable Change Index (RCI) method. Means not sharing a common subscript were significantly different (p < .05) in posthoc comparisons using the Tukey-b test.