Published in final edited form as:

J Abnorm Child Psychol. 2013 November ; 41(8): 1267–1277. doi:10.1007/s10802-013-9756-6.

Comparisons across depression assessment instruments in adolescence and young adulthood: An Item Response Theory study using two linking methods

Thomas M. Olino^{1,‡}, Lan Yu¹, Dana L. McMakin¹, Erika E. Forbes¹, John R. Seeley², Peter M. Lewinsohn², and Paul A. Pilkonis¹

¹University of Pittsburgh

²Oregon Research Institute

Abstract

Item response theory (IRT) methods allow for comparing the utility of instruments based on the range and precision of severity assessed by each instrument. As adolescents and young adults can display rapid increases in depressive symptoms, there is a crucial need to sensitively assess mild elevations of symptoms (as an index of initial risk) and moderate-severe symptoms (as an indicator of treatment disposition). We compare the information assessed by the Beck Depression Inventory (BDI) to the newly developed Patient Reported Outcome Measurement Information System – Depression measure (PROMIS-Depression), and the Center for Epidemiologic Studies – Depression (CES-D) scale. The present work is based on data from two fully independent samples of community adolescents and young adults. One sample completed the BDI and CES-D ($n = 1482$) and the second sample ($n = 673$) completed the PROMIS-Depression measure and the CES-D. Using two different IRT-based linking methods, (1) equating based on common items and (2) concurrent calibration methods, analyses revealed that the PROMIS-Depression measure assessed information over the widest range of depressive severity with greatest measurement precision relative to the other instruments. This was true for both the 28-item and 8-item versions of the PROMIS-Depression measure. Findings suggest that the PROMIS-Depression measure assessed depression severity with greatest precision and over the widest severity range of the assessed instruments. However, future work is necessary to demonstrate that the PROMIS-Depression measure has reliable associations with external criteria and is sensitive to treatment response.

Keywords

adolescent depression assessment; item response theory; psychometrics

Selection of instruments to assess depressive symptomatology throughout adolescence and into young adulthood is crucial. During adolescence, the incidence of depression increases (Hankin, et al., 1998; Lewinsohn, Clarke, Seeley, & Rohde, 1994) and symptom severity is often a factor that influences the level of care and treatment modalities employed (Fournier, et al., 2010). In addition, it is important to assess modest elevations in symptom severity to identify youth with increasing levels of symptoms that may serve as a risk for escalation to full threshold depressive disorders in the future (Klein, Shankman, Lewinsohn, & Seeley, 2009). Historically, different measures are used to assess symptom levels when severity is

[‡]Please direct correspondence to: Thomas M. Olino, Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, 15213. Phone: (412)383-5434; fax (412) 383-5426; olinotm@upmc.edu.

The authors have no other financial disclosures. The authors report no conflicts of interest.

expected to be high or low in the sample (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961; Radloff, 1977). Thus, there are few instruments that can be effectively used to identify individuals with modest versus clinically significant symptom levels (Olino, et al., 2012). Classical psychometric methods focus on internal consistency; test-retest reliability; predictive validity; positive and negative predictive power; and sensitivity and specificity, with many instruments demonstrating similar characteristics in adolescents and adults (Klein, Dougherty, & Olino, 2005; Roberts, Lewinsohn, & Seeley, 1991). In addition, the development of norms for instruments using classical test theory relies on sample specific information that defines rank order standing (i.e., percentile ranks). Thus, these measures provide modest guidance about how these instruments differ with respect to the range of severity that they assess independent of a particular sample. Modern measurement methods, particularly item response theory (IRT; Embretson & Reise, 2000), provide significant leverage towards characterizing the range of severity assessed by specific instruments and their included items.

IRT is a set of modern psychometric techniques that model levels of a dimension (i.e., latent trait) as a function of item characteristics that vary from item to item, which departs from assumptions of uniform item functioning from classical psychometric theory. A critical and inherent requirement for implementing IRT is that all items assess the same underlying construct. As each individual set of items aligns along the same dimension, individual items across multiple measures may be compared on the same metric.

IRT methods define where on the dimension (i.e., theta level) an item provides the most information and over what range of severity the item provides information using specific item parameters: difficulty and discrimination (Embretson & Reise, 2000). In the context of depression, items measuring suicidality may provide information at higher severity levels than items measuring depressed mood. Difficulty and discrimination parameters for items can be integrated graphically to provide item information functions that display the range of the dimension assessed by that particular item. Difficulty parameters indicate the specific level of severity level whereby individuals have a 50% chance of endorsing the item and discrimination parameters indicate the strength of association between the item and latent factor. Parameters from multiple items or full item sets from instruments may be aggregated to provide total test information, which yields evidence about the range of the dimension assessed by the instrument, what range of the dimension is best assessed by the instrument, and, by extension, how different instruments provide similar or complementary information.

Many studies of adolescents have utilized the depression instruments developed with adults, including the Beck Depression Inventory (BDI; Beck et al., 1961) and the Center for Epidemiologic Studies-Depression scale (CES-D; Radloff, 1977). As the BDI was developed for clinical monitoring purposes among patient populations, the intention was to assess severity at higher depression levels. Additional instruments (e.g., Center for Epidemiologic Studies-Depression [CES-D]; Radloff, 1977) were subsequently developed to assess symptomatology more broadly for use in epidemiological studies. However, there have been few attempts, almost exclusively in adult samples (for exceptions, see Carmody, et al., 2006; Santor, Zuroff, Ramsay, Cervantes, & Palacios, 1995; Uher, et al., 2008), to empirically compare whether instruments assess depression severity over similar or different ranges. Using IRT methods, Santor et al. (1995) found that the CES-D provided better discriminability of depressive severity at lower levels of depression in a college student sample and a depressed, adult outpatient sample than the BDI. Only one previous study has used these methods to compare measures in an adolescent sample. Using IRT concurrent calibration methods, Olino et al. (2012) built on this work and found that both the BDI and CES-D assessed severity along a wide range, however, the CES-D assessed severity better

than the BDI at lower levels of depression in a community sample of adolescents (mean age = 16).

More recently, instruments have been developed using modern psychometric methods (i.e., item response theory) in an attempt to maximize the range of severity assessed, while minimizing participant/client burden for both adult and children in a number of physical and mental health domains (DeWitt, et al., 2011; Irwin, et al., 2012; Irwin, et al., 2010; Pilkonis, et al., 2011). However, comparisons of instruments developed using traditional and modern psychometric methods are only recently being conducted.

The National Institutes of Health initiated a Patient Reported Outcomes Measurement Information System (PROMIS®) network that used modern psychometric methods to develop instruments to assess multiple domains of psychological and physical health. The PROMIS network used IRT principles to develop and refine items that reflect a unidimensional, latent depression dimension (Pilkonis et al., 2011) that ultimately resulted in the development of computerized adaptive assessments (Gibbons, et al., 2012). PROMIS collected data from individuals in community, psychiatric, and physical health settings to maximize generalizability and utility of the developed instrument. In the development of the depression measure for adults, the network surveyed individual items from publically available instruments. Field testing yielded a 28-item bank that was intended to be used for computerized adaptive testing, to minimize administration times, and a short, 8-item static measure (Pilkonis et al., 2011). In the development sample, the PROMIS-Depression measure was administered along with the CES-D and concurrent calibration IRT analyses were used to compare the range of severity assessed by the short-form and long-form PROMIS-Depression measures and the CES-D. This comparison revealed that the short-form PROMIS measure and CES-D provided information along similar ranges of depression severity. However, the long-form PROMIS-Depression measure provided more information and information along a wider range of depression than the short-form and the CES-D.

In addition to direct comparison using calibration methods utilized in the aforementioned studies (i.e., Olino et al., 2012; Pilkonis et al., 2011), additional IRT approaches are available that facilitate comparing measures that are administered to independent samples. As principles underlying IRT methods argue that parameters recovered from analyses are properties of the items, rather than being functions of specific samples, these item banks can serve an instrumental function by anchoring common items from independent samples to the same parameter values. Non-common items can be linked across different samples based on relationships to the common item set. To our knowledge, this methodology has not been employed in the area of depression assessment at all, let alone with adolescents and/or young adults.

The present study examines the IRT-based functioning of the PROMIS-Depression short- and long-form, CES-D, and BDI. Here, we present comparisons of these instruments by relying on independent samples using two different IRT linking methods. Based on previous results presented by Pilkonis et al. (2011) and Olino et al. (2012) we anticipate that PROMIS-Depression will provide information over a wider severity range and will provide greater information than the CES-D and BDI across both analytic methods.

Methods

Data for the present study came from two independent samples. We present sample information for each sample and the assessment methods for each.

Oregon Adolescent Depression Project

(OADP; Lewinsohn, Hops, Roberts, Seeley, & Andrews, 1993). Participants were randomly selected from nine high schools in western Oregon from 1987–1989. The participation rate was 61% and comparisons between the 1980 census data and the recruited sample revealed no significant differences on gender, ethnicity, or parental education level. Before study related procedures commenced, signed informed consent was received by study staff. The OADP began with a total of 1,709 adolescents (ages 13.89–20.01; mean age 16.56, $SD = 1.19$; 52.1% [$n = 891$] female; 91.1% [$n = 1557$] Caucasian, 1.0% [$n = 17$] African-American, 2.5% [$n = 42$] Hispanic, 2.2% [$n = 37$] Indian, 1.9% [$n = 33$] Asian, 1.3% [$n = 23$] Other). These participants served as the sample reported on in Olino et al. (2012). The present report focuses on the participants who completed the second assessment that occurred approximately one year after the initial assessment. For this wave of data collection, 88% ($n = 1,507$) of the adolescents returned for a second evaluation (T2; an approximate 50% participation rate of the sample that was initially asked to participate). Several checks were conducted to assess the representativeness of the sample (see Lewinsohn et al., 1993 for details). Briefly, adolescents who did and did not participate in the T1 assessment were similar on intactness of the family and family size, but participants were more likely to be female, in a higher grade, and their parental socioeconomic status was lower than non-participants. Of the adolescents who participated in the T1 assessment, those who also participated at T2 were more likely to be female, came from larger families, had a higher parental SES, and were less likely to have had a T1 diagnosis of a disruptive behavior disorder, and, for males only, a T1 substance use disorder, compared to those who did not participate at T2. However, among T1 participants, those who did and did not participate at T2 were similar on race, grade level, and all other T1 current and lifetime diagnoses.

All participants were asked to complete the 20 item CES-D (Radloff, 1977) and the 21 item BDI (Beck, Steer, & Carbin, 1988). Of the 1507 total participants at T₂, 1482 (98.3%) of participants completed some responses on either the CES-D or BDI. This sample is the focus of the present study. The mean age of this sample was 17.6 years ($SD = 1.20$; range 14.90–21.07 years); 53.8% female; and 91.9% were Caucasian, 0.5% African American, 1.5% Asian, and 5.2% Other.

The CES-D and BDI are two commonly used self-report instruments to assess depression in both youth and adult samples (Klein et al., 2005; Lipsman & Lozano, 2011). They have both demonstrated high levels of internal consistency and test-retest stability. Responses for the 20 items on the CES-D are frequency based, with four response options that range from 0 (*rarely/none*) of the time to 3 (*most/all of the time*). Responses for the 21 item BDI are severity based and include four different severity levels per item that are individualized for each items. The BDI and CES-D assess symptoms as they have been experienced in the past week.

Personality, Anhedonia, and Depression (PAD)

PAD is a psychometric study of the inter-relationships among measures assessing approach motivational system constructs. The study included 698 undergraduate students from the psychology subject pool from the University of Pittsburgh. To maintain comparability of the samples in late adolescence and extending into young adulthood, only PAD participants younger than 22 years ($n = 673$) were retained in the present sample. The mean age of this sample was 19.09 years ($SD = 0.81$; range 18.00–21.93 years); 64.5% were female; and 79.8% were Caucasian, 6.6% were African American, 8.7% were Asian, and 5.2% were other races. Before study related procedures commenced, signed informed consent was received by study staff.

Participants completed the CES-D and the PROMIS-Depression scale (Pilkonis et al., 2011). Responses for the 28 items on the PROMIS-Depression scale are endorsed on a 5-point frequency-based Likert scale with five response options that range from 0 (*Never*) to 4 (*Always*) within the past week. In addition to the full 28-item measure, the original authors identified a subset of 8-items that could be used as an abbreviated measure.

Analytic Approach—IRT refers to a family of models in which the probability of correct item response is modeled as a function of latent trait θ and one or more item parameters (Lord, 1980). The most commonly used IRT model for polytomous item responses is the Graded Response Model (GRM; Samejima, 1969). GRM provides one discrimination parameter and $n-1$ threshold parameters for each item, where n is the number of response categories. The probability of endorsing a specific response option takes the form of multiple distributions along the θ dimension. Discrimination values greater than 1 are generally accepted as capturing acceptable amounts of information (Baker, 2001). IRT-based instrument development and evaluation typically is motivated to measure levels of severity at levels both much below and much higher than the mean level of the trait. Thus, difficulty parameters will range in values with some items providing information at lower and others at higher levels. The precise shape of an item characteristic curve (ICC) depends on the values of item parameters, with the difficulty parameters indicating the location of equal endorsement probabilities for adjacent responses and the difficulty parameter indicating the rate of change (with respect to the underlying dimension) of probability of response endorsement. An ICC can be transformed into an *item information curve*, indicating the amount of psychometric information an item contains at all points along the scale. Multiple item information curves can be aggregated together to form a *test information curve* (TIF), which indicates the amount of information a whole test contains at all points along the scale. More information reflects greater measurement precision, or reliability. Therefore, an important feature of IRT models is that reliability is described as a function conditional on values of θ . This is in contrast to classical test theory that assumes uniform reliability across all levels of scores. Another advantage of IRT is that individuals' estimates are independent of the number of items or the specific items used for testing.

Dimensionality: IRT models are based on a number of assumptions. One important assumption is that the underlying being measured is unidimensional. Previous publications investigating the measures in the current study (e.g., Olino et al., 2012; Pilkonis et al., 2011) found supporting evidence for unidimensionality using both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) for all individual and aggregate instruments. As no participants completed both the PROMIS-Depression and the BDI items across samples, we cannot provide an examination of unidimensionality for the full item pool. However, we examined EFA and CFA models for the CES-D, which was common to both samples, to provide support for the common items being unidimensional across samples.

For EFA analyses, we relied on the ratio of the first to second eigenvalues to support unidimensionality. A generally accepted recommendation is that unidimensionality is supported when the ratio of the first to second eigenvalue exceeds 4 (Lord, 1980; Reise & Waller, 1990); however, more recently an eigenvalue ratio exceeding 3 is supportive of unidimensionality (Embretson & Reise, 2000). We rely on the more conservative eigenvalue ratio of 4 for the present work. For CFA models, we assessed absolute fit of the confirmatory models using global fit indices, including the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). For the CFI and TLI, we used the conventional cutoff values .90 or greater for acceptable fit, and .95 or greater for good fit. RMSEA values between .05 and .10 represent an acceptable fit (Steiger, 1990), whereas values less than .05 indicate a good fit (McDonald & Ho, 2002).

Equating based on common items and concurrent calibration: We calibrated all items using Multilog 7.03 (Thissen, Chen, & Bock, 2003). The first analytic method, equating based on common items, linked the two independent samples through common CES-D items. This method used the underlying construct of PROMIS-Depression items and put both CES-D and BDI items, in separate steps, on the same underlying construct of PROMIS-Depression items. This method included three steps. Step 1 calibrated the PROMIS-Depression items using the PAD sample. This created the item bank information for the subsequent linking analysis. In step 2, the CES-D items were calibrated with the PROMIS-Depression item parameters fixed (i.e., parameter estimates obtained from step 1 for the PROMIS sample). Step 3 calibrated the BDI items using the OADP sample with the CES-D item parameters fixed to the values obtained from step 2. Thus, across the two independent samples, the item banking process facilitated a comparison across the CES-D, BDI, and PROMIS-Depression scales.

The second analytic method utilized concurrent calibration methods. This method assumed the underlying construct is the combined PROMIS-Depression items, BDI items, and CES-D items. Thus, all items from both samples were included in the analysis and absent items across samples were treated as missing by design (i.e., missing completely at random). Item parameters were estimated within a single run. In order to compare the results from the two methods, intra-class correlation coefficients were estimated as a means of comparing recovered item parameters across each analytic method. Intra-class correlation coefficients were estimated using two-way random effects with absolute agreement to incorporate rank- and mean-level similarity in estimates across analytic strategy (Shrout & Fleiss, 1979).

Results

Within the OADP sample, the median and mean scores (and standard deviation) for the CES-D and BDI were 12.00 and 14.26 (10.44), and 3.00 and 5.41 (7.14), respectively. A small number of participants (3.0% [$n = 45$]) scored a zero on the CES-D and a larger number (24.4% [$n = 362$]) scored a zero on the BDI. Within the PAD sample, the median and mean (and standard deviation) for the CES-D and PROMIS-Depression were 12.00 and 13.87 (9.53), and 24.00 and 26.98 (19.78), respectively. A modest number of participants (0.4% [$n = 3$]) scored a zero on the CES-D and a modest number of participants scored a zero on the PROMIS-Depression measure (1.2% [$n = 8$]). The samples did not significantly differ on mean CES-D scores, $t(2149) = .83$, $p = .40$. However, the OADP sample had a higher proportion of female participants (64.5% vs. 53.8%, $\chi^2(1) = 21.67$, $p < .001$) and a lower proportion of racial minority participants (8.4% vs. 20.2%, $\chi^2(1) = 21.67$, $p < .001$) than the PAD sample. Table 1 provides percentile ranks corresponding to raw scores for each instrument.

Assessing Dimensionality

An EFA and CFA were estimated on the CES-D using Mplus 6.12 (Muthén & Muthén, 1998–2010) using the WLSMV estimator. Each of the samples were randomly assigned to either contributing to EFA or CFA models (OADP $n = 763$ and 719, respectively; PAD $n = 331$ and 342, respectively) using a random number generator with binary outcomes. Table 2 presents relevant model fit information from the EFA and CFA analyses for each sample and for the CES-D across both samples.

For the EFA of the CES-D in the OADP sample, the ratio of the first to second eigenvalues exceeded the rule of thumb value of 4 (Lord, 1980; Reise & Waller, 1990) and was consistent with unidimensionality. For the CFA of the CES-D in the OADP sample, all indices of model fit were acceptable. For the EFA of the BDI in the OADP sample, the ratio of the first to second eigenvalues exceeded the recommended value of 4 and was consistent

with unidimensionality. For the CFA of the BDI in the OADP sample, all indices of model fit were excellent. For the EFA of the CES-D in the PAD sample, the ratio of the first to second eigenvalues exceeded the recommended value of 4 and was consistent with unidimensionality. For the CFA of the CES-D in the PAD sample, across all indices model fit was acceptable. For the EFA of the PROMIS-Depression scales in the PAD sample, the ratio of the first to second eigenvalues exceeded the recommended value of 4 and was consistent with unidimensionality. For the CFA of the PROMIS-Depression scales in the PAD sample, across all indices model fit was good. Lastly, we assessed whether the CES-D was unidimensional when including individuals from both samples. Thus, we examined EFA and CFA models on randomly divided portions of the total sample. Across both samples, the EFA revealed an eigenvalue ratio exceeding three, and the CFA found that a one-factor model was an acceptable fit to the data. Thus, each individual instrument within each sample demonstrated sufficient unidimensionality to proceed with IRT-based analytic methods.

IRT Analyses

Equating based on common items—First, the PROMIS-Depression measure was calibrated individually. The discrimination parameters ranged from .98 to 3.37 and first, second, third, and fourth difficulty parameters ranged from −1.36 to 1.32, −.23 to 2.23, 1.00 to 2.74, and 2.35 to 4.05, respectively. Second, we calibrated the PROMIS-Depression measure, using the item bank from the initial analysis, with the CES-D within the PAD sample. For the CES-D, the discrimination parameters ranged from .70 to 2.74 and first, second, and third difficulty parameters ranged from −1.01 to 1.08, .58 to 2.08, and 1.72 to 4.03, respectively. Lastly, the item parameters for the CES-D recovered from the PAD sample were used as an item bank to calibrate the BDI items in the OADP sample. For the BDI items, discrimination parameters ranged from .57 to 2.43 and first, second, and third difficulty parameters ranged from .58 to 3.19, 2.25 to 5.56, and 2.85 to 7.44, respectively. Individual parameter estimates are displayed in Table 2.

IRT Concurrent Calibration: This analysis calibrated all items from the BDI, PROMIS-Depression, and CES-D simultaneously. Similar to the initial analytic strategy, the PROMIS-Depression measure's discrimination parameters ranged from 1.07 to 3.96 and first, second, third, and fourth difficulty parameters ranged from −1.22 to 1.20, −.20 to 2.05, 0.88 to 2.53, and 2.09 to 3.73, respectively. For the CES-D, the discrimination parameters ranged from 0.73 to 2.40 and first, second, and third difficulty parameters ranged from −1.10 to 1.19, 0.59 to 2.41, and 1.98 to 4.26, respectively. For the BDI items, discrimination parameters ranged from 0.54 to 2.39 and first, second, and third difficulty parameters ranged from 0.59 to 3.23, 2.28 to 5.64, and 2.90 to 7.55, respectively.

Test Information: Figure 1 displays the test information function (TIF) for the full PROMIS-Depression, short PROMIS-Depression, CES-D, and BDI using the parameters from the common item linking method (Top Panel) and from the concurrent calibration method (Bottom Panel). As a means of standardizing our description of instrument functioning, we use test information value of approximately 10 to reflect reliability of .90 from classical test theory (Yu, et al., 2011). Using this metric as a guide and based on the common item linking method (Figure 1, Top Panel), the full PROMIS-Depression measure assessed information well from approximately 1.50 SD below through 4.00 SD above the mean of depression severity. The short PROMIS-Depression measure assessed information well from approximately 0.50 SD below the mean through 3.25 SD above the mean of depression severity. The CES-D assessed information well from approximately 0.50 SD below the mean through 2.50 SD above the mean of depression severity. Finally, the BDI assessed information well from approximately 0.50 SD above the mean through 4.00 SD

above the mean of depression severity. Across measures, the full PROMIS-Depression measure assessed information across the widest range with most measurement precision. While the full PROMIS measure provided abundantly more information than any other measure, the eight item PROMIS measure provided information at a wider range and with generally at least comparable precision to that of the BDI and CES-D.

A generally similar pattern of results is displayed when using the concurrent calibration methods (Figure 1, Bottom Panel). Again, using information of 10 being similar to reliability of .90, the full PROMIS-Depression measure assessed information well from approximately 1.50 SD below through 4.00 SD above the mean of depression severity. The short PROMIS-Depression measure assessed information well from approximately 0.75 SD below the mean through 3.25 SD above the mean of depression severity. The CES-D assessed information well from approximately 0.25 SD below the mean through 2.75 SD above the mean of depression severity. Finally, the BDI assessed information well from approximately 0.50 SD *above* the mean through 4.00 SD above the mean of depression severity. Across measures, the full PROMIS-Depression measure assessed information across the widest range with most measurement precision. Again, whereas the full PROMIS measure provided much more information than any other measure, the eight item PROMIS measure provided information at a wider range and with generally at least comparable precision to that of the BDI and CES-D.

Comparison of Parameters Across Analytic Method

Intra-class correlation coefficients were estimated as a means of quantifying the consistency in the results across analytic methods. Associations were estimated for specific parameters (i.e., discrimination, first difficulty, second difficulty, third difficulty, and fourth difficulty [for the PROMIS-Depression items, only]) for each instrument separately. Similar to the consistency in the reported range of specific parameter values, there was excellent agreement for difficulty and discrimination parameters across all instruments (all intra-class correlation coefficients $\geq .85$).

Discussion

The incidence of depression increases dramatically during adolescence (Hankin et al., 1998; Lewinsohn et al., 1993); thus, there is a pressing need for recognizing mild elevations of symptoms for identifying those at risk and for recognizing moderate-severe symptom levels for guiding appropriate treatment recommendations. However, few investigations have used modern measurement techniques, specifically IRT, to compare and understand how instruments provide information at different or similar levels of severity. The present study used data from two independent samples consisting of older adolescents and young adults to evaluate the effective ranges of depression severity for the BDI, CES-D, and newly developed PROMIS-Depression measure.

Findings reported here are consistent with previous work in distinct samples. In an adult sample, Pilkonis et al. (2011) reported that the PROMIS-Depression measure provided information over a wider severity range than the CES-D. In a younger adolescent sample, Olino et al. (2012) reported that the BDI and CES-D assessed information at largely overlapping ranges, with the BDI performing better at modestly higher levels of severity. The current results extend these comparisons by finding that the PROMIS-Depression measure assesses depressive severity along a broader range of severity, with greater precision, than the BDI. The full 28-item PROMIS instrument provided more than double the total psychometric information provided the CES-D or the BDI. Impressively, the 8-item PROMIS short-form assessed information well over a wider range than the BDI and CES-D items. These results suggest that the PROMIS measure should provide valid information

across multiple settings for an array of purposes, including those settings where lower levels of depressive symptomatology are expected (e.g., epidemiologic studies) and settings where symptom levels are expected to be much higher (e.g., clinical trials). The context of clinical trials is particularly important as instruments that are insensitive (i.e., do not assess information) at lower trait levels may obscure important variability at lower levels of severity observed at the end stages of treatment and erroneously identify treatment plateaus. Whereas previous work with the PROMIS-Depression items have been utilized with adults, the present study finds that these items work well with older adolescents.

In contrast to these highlighted differences, there were also levels of severity that provided highly reliable information across all instruments. In particular, highly reliable information was available from approximately 0.50 SD above the mean through approximately 2.50 SD above the mean. Thus, for assessing individuals expected to be within this range of severity, items from any of the evaluated instruments would be reasonable.

These substantive comparisons were made feasible by utilizing an important feature of IRT: linking methods. These methods permit comparison of items from independent samples when the same latent construct is being assessed. This practice would facilitate the broader comparisons of instruments based on these item parameters. For example, future work could compare the functioning of additional measures (e.g., the Patient Health Questionnaire-9; Kroenke, Spitzer, & Williams, 2001; Quick Inventory of Depressive Symptomatology; Rush et al., 2003) that are administered along with any of the instruments included in the current study using item banking procedures. Thus, a network of IRT-based instrument functioning can be established. While this has been a long described extension of IRT methods, this has not been fully realized in the literature, particularly in the area of clinical assessment. Impressively, results across two analytic approaches yielded highly similar findings. The item parameters recovered from the banking and linking method were highly similar to those recovered from the concurrent calibration approach. This is particularly interesting in light of the way that the equating procedure was based on initial calibration of the PROMIS-Depression items, which placed all other items on the underlying construct as defined by this instrument. Thus, the convergence across methods suggests that the underlying depression construct is remarkably similar across the items (and measures) sampled.

While the visual representation of the findings is particularly striking, these results should be considered in light of some methodological issues. The amount of information and range of severity assessed is influenced by the number of items and the number of response options per item. The full PROMIS-Depression measure includes more items than any other instrument included and the items included more response options than items from any other measure. Thus, these differences could account for the broad conclusion that the PROMIS-Depression measure assesses information over the widest range with greatest precision relative to other instruments included here. However, the short-form, 8-item, PROMIS-Depression measure provides more information than any other measure, suggesting that even the 8-item measure could be preferred. There are also differences between the response scales for the PROMIS-Depression and CES-D relative to the BDI. The PROMIS-Depression and CES-D ask for frequency based ratings whereas the BDI asks for intensity based ratings. Thus, some differences in results may be a result of these differences, which would suggest that these responses provide information on different underlying latent dimensions. Previous work (Olino et al., 2012), however, found that the BDI and CES-D, along with items from a diagnostic interview, tapped the same underlying dimension.

The present study benefitted from two reasonably large samples and implementation of sophisticated modern measurement methods to specifically address whether depression measures assess severity over the same or different levels. However, the study was limited

by a number of factors. First, we focused on depression assessment instruments that were originally developed for use with adults. It will be important to examine whether the underlying depression severity dimension for these instruments can be mapped onto the same underlying dimension as assessed by instruments developed specifically for adolescents (e.g., MFQ & CDI, respectively, Angold, Costello, Messer, & Pickles, 1995; Kovacs, 1992). This could be greatly benefitted by utilizing item banking procedures employed here. Second, while we combined data from two independent samples, one sample included adolescents who were largely attending high school; the other sample included older adolescents who were attending university. Thus, there may be some developmental differences between these samples that may limit comparability. However, no significant differences in mean CES-D scores were found and we found similar factor loadings results across EFA and CFA methods in both samples, suggesting that the information provided in each sample would be similar. Third, although both samples were predominantly Caucasian, the racial composition of the samples differed. As there is evidence about the role of race in manifestations of depression, there may be impacts on functioning of self-report items (Steele, et al., 2006; Twenge & Nolen-Hoeksema, 2002). However, our sample size was too modest to examine this issue. Thus, future studies should be designed to directly test this question.

The present study compared the BDI, CES-D, and PROMIS-Depression scales using IRT and found that the PROMIS-Depression scale provided information over the greatest range of depressive severity with the greatest precision of measurement. This suggests that this measure is a robust instrument for studies and practice spanning epidemiology and clinical trials, whereas other instruments are better suited for a more restrictive set of goals. Future work should examine how these adult measures collected on adolescents compare with adolescent-oriented measures collected on adolescent samples.

Acknowledgments

The present work was supported by K01 MH092603 (TMO) and R01 MH40501 (PML).

References

- Angold A, Costello EJ, Messer SC, Pickles A. Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *International Journal of Methods in Psychiatric Research*. 1995; 5:237–249.
- Baker, FB. The basics of item response theory. ERIC; 2001.
- Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*. 1988; 8:77–100.
- 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]
- Carmody TJ, Rush A, Bernstein IH, Brannan S, Husain MM, Trivedi MH. Making clinicians lives easier: Guidance on use of the QIDS self-report in place of the MADRS. *Journal of Affective Disorders*. 2006; 95:115–118. [PubMed: 16707162]
- DeWitt EM, Stucky BD, Thissen D, Irwin DE, Langer M, Varni JW, et al. Construction of the eight-item patient-reported outcomes measurement information system pediatric physical function scales: built using item response theory. *Journal of Clinical Epidemiology*. 2011; 64:794–804. [PubMed: 21292444]
- Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
- Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, et al. Antidepressant drug effects and depression severity. *Journal of the American Medical Association*. 2010; 303:47–53. [PubMed: 20051569]

- Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, et al. Development of a computerized adaptive test for depression. *Archives of General Psychiatry*. 2012; 69:1104–1112. [PubMed: 23117634]
- Hankin BL, Abramson LY, Moffitt TE, Silva PA, McGee R, Angell KE. Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology*. 1998; 107:128–140. [PubMed: 9505045]
- Irwin DE, Gross HE, Stucky BD, Thissen D, DeWitt EM, Lai JS, et al. Development of six PROMIS pediatrics proxy-report item banks. *Health and Quality of Life Outcomes*. 2012; 10:22. [PubMed: 22357192]
- Irwin DE, Stucky B, Langer MM, Thissen D, DeWitt EM, Lai JS, et al. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*. 2010; 19:595–607. [PubMed: 20213516]
- Klein DN, Dougherty LR, Olino TM. Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*. 2005; 34:412–432. [PubMed: 16026212]
- Klein DN, Shankman SA, Lewinsohn PM, Seeley JR. Subthreshold depressive disorder in adolescents: Predictors of escalation to full-syndrome depressive disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2009; 48:703–710. [PubMed: 19465876]
- Kovacs, M. Children's depression inventory. North Tonawanda, N.Y.: Multi-Health System; 1992.
- Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. *Journal of General Internal Medicine*. 2001; 16:606–613. [PubMed: 11556941]
- Lewinsohn PM, Clarke GN, Seeley JR, Rohde P. Major depression in community adolescents: Age at onset, episode duration, and time to recurrence. *Journal of the American Academy of Child & Adolescent Psychiatry*. 1994; 33:809–818. [PubMed: 7598758]
- Lewinsohn PM, Hops H, Roberts RE, Seeley JR, Andrews JA. Adolescent psychopathology: I. Prevalence and incidence of depression and other DSM-III—R disorders in high school students. *Journal of Abnormal Psychology*. 1993; 102:133–144. [PubMed: 8436689]
- Lipsman N, Lozano AM. The most cited works in major depression: The 'Citation classics'. *Journal of Affective Disorders*. 2011; 134:39–44. [PubMed: 21652082]
- Lord, FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum; 1980.
- McDonald RP, Ho MHR. Principles and practice in reporting structural equation analyses. *Psychological Methods*. 2002; 7:64–82. [PubMed: 11928891]
- Muthén, LK.; Muthén, BO. Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén; 1998–2010.
- Olino TM, Yu L, Klein DN, Rohde P, Seeley JR, Pilkonis PA, et al. Measuring depression using item response theory: An examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research*. 2012; 21:76–85. [PubMed: 22290656]
- 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. The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*. 1977; 1:385–401.
- Reise SP, Waller NG. Fitting the two-parameter model to personality data: The parameterization of the Multidimensional Personality Questionnaire. *Applied Psychological Measurement*. 1990; 14:45–58.
- Roberts RE, Lewinsohn PM, Seeley JR. Screening for adolescent depression: a comparison of depression scales. *Journal of American Academy of Child & Adolescent Psychiatry*. 1991; 30:58–66.
- Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*. 2003; 54:573–583. [PubMed: 12946886]

- Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969; No. 17
- Santor DA, Zuroff DC, Ramsay JO, Cervantes P, Palacios J. Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment*. 1995; 7:131–139.
- Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. [PubMed: 18839484]
- Steele RG, Little TD, Ilardi SS, Forehand R, Brody GH, Hunter HL. A confirmatory comparison of the factor structure of the Children's Depression Inventory between European American and African American youth. *Journal of Child and Family Studies*. 2006; 15:773–788.
- Steiger JH. Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*. 1990; 25:173–180.
- Thissen, D.; Chen, W-H.; Bock, RD. MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.; 2003.
- Twenge JM, Nolen-Hoeksema S. Age, gender, race, socioeconomic status, and birth cohort difference on the children's depression inventory: A meta-analysis. *Journal of abnormal psychology*. 2002; 111:578. [PubMed: 12428771]
- Uher R, Farmer A, Maier W, Rietschel M, Hauser J, Marusic A, et al. Measuring depression: Comparison and integration of three scales in the GENDEP study. *Psychological Medicine*. 2008; 38:289–300. [PubMed: 17922940]
- Yu L, Buysse DJ, Germain A, Moul DE, Stover A, Dodds NE, et al. Development of Short Forms From the PROMIS™ Sleep Disturbance and Sleep-Related Impairment Item Banks. *Behavioral Sleep Medicine*. 2011; 10:6–24. [PubMed: 22250775]

Figure 1.

CES-D: Center for Epidemiologic Studies – Depression; BDI: Beck Depression Inventory. Top panel: Test information plot for parameters recovered from the common item calibration methods. Final parameters were estimated using a sequential linking process that began by calibrating the PROMIS-Depression items. Then the CES-D items were calibrated with the PROMIS-Depression items fixed at the previously recovered values. Finally, the BDI items were calibrated with the CES-D items fixed at the previously recovered values. Thus, the CES-D items were used as common items to link the samples. Bottom panel: Test information plot for parameters recovered from the concurrent calibration methods. Final

model parameters were estimated by calibrating all items across both samples simultaneously.

Table 1

Raw scores for the CES-D, BDI, and PROMIS-Depression corresponding to percentile ranks.

Percentile	CES-D	BDI	PROMIS-Depression
10	3	0	33
20	6	0	38
30	7	1	42
40	9	2	46
50	12	3	51
60	14	4	56
70	18	6	62
80	22	9	70
90	29	14	84

Note: CES-D $n = 2151$; BDI $n = 1476$; and PROMIS-Depression $n = 673$. For an interpretation example, youth scoring a 7 on the CES-D scale, a 1 on the BDI, and a 42 on the PROMIS-Depression measure all scored at the 30th percentile.

Table 2
Results of Unidimensionality Analyses for the BDI, CES-D, PROMIS-Depression.

	First Eigenvalue	EFA Second Eigenvalue	Ratio	CFI	CFA TLI	RMSEA
OADP						
BDI	10.28	1.32	7.79	.98	.98	.04 (.03–.04)
CES-D	8.89	1.45	6.13	.92	.90	.09 (.08–.09)
PAD						
CES-D	9.14	1.53	5.97	.92	.91	.09 (.09–.10)
PROMIS	16.98	1.39	12.22	.97	.97	.07 (.06–.07)
Combined						
CES-D	8.82	1.39	6.35	.91	.90	.09 (.08–.09)

Note: OADP = Oregon Adolescent Depression Project sample; PAD = Personality, Anhedonia, Depression sample; BDI = Beck Depression Inventory; CES-D = Center for Epidemiologic Studies-Depression scale; PROMIS = PROMIS Depression Scale.

Table 3

Item parameters from equating and concurrent calibration methods

Item		Equating					Concurrent Calibration				
		a	b1	b2	b3	b4	a	b1	b2	b3	b4
PROMIS 1	I felt worthless	2.44	-0.11	0.81	1.97	3.21	2.73	-0.10	0.72	1.77	2.94
PROMIS 2	I felt that I had nothing to look forward to	1.88	-0.10	1.03	2.09	3.25	2.04	-0.09	0.93	1.90	2.98
PROMIS 3	I felt helpless	2.35	-0.35	0.69	1.74	3.13	2.55	-0.31	0.62	1.58	2.90
PROMIS 4	I withdrew from other people	1.81	-0.78	0.37	1.75	3.78	2.04	-0.69	0.33	1.57	3.42
PROMIS 5	I felt that nothing could cheer me up	2.74	0.06	0.91	1.85	3.01	3.00	0.05	0.81	1.69	2.77
PROMIS 6	I felt that I was not as good as other people	1.87	-0.55	0.37	1.56	2.87	2.07	-0.48	0.33	1.41	2.61
PROMIS 7	I felt sad	2.30	-1.36	-0.14	1.37	2.92	2.67	-1.17	-0.12	1.21	2.63
PROMIS 8	I felt that I wanted to give up on everything	2.22	0.06	0.94	1.92	3.08	2.42	0.05	0.85	1.74	2.83
PROMIS 9	I felt that I was to blame for things	1.79	-0.19	0.93	2.33	3.31	1.91	-0.17	0.85	2.15	3.07
PROMIS 10	I felt like a failure	2.39	-0.03	0.88	1.97	2.76	2.52	-0.03	0.80	1.81	2.55
PROMIS 11	I had trouble feeling close to people	1.68	-0.17	0.84	1.83	3.03	1.84	-0.15	0.76	1.66	2.76
PROMIS 12	I felt disappointed in myself	2.01	-1.05	0.13	1.54	2.44	2.25	-0.92	0.12	1.37	2.20
PROMIS 13	I felt that I was not needed	2.33	0.00	0.85	1.91	2.89	2.52	0.00	0.77	1.74	2.65
PROMIS 14	I felt lonely	2.28	-0.55	0.32	1.31	2.43	2.63	-0.48	0.28	1.16	2.17
PROMIS 15	I felt depressed	3.37	-0.13	0.69	1.56	2.49	3.96	-0.11	0.60	1.39	2.26
PROMIS 16	I had trouble making decisions	0.98	-1.36	0.40	2.37	3.82	1.07	-1.22	0.37	2.16	3.48
PROMIS 17	I felt discouraged about the future	1.99	-0.69	0.45	1.44	2.50	2.25	-0.60	0.40	1.29	2.25
PROMIS 18	I found that things in my life were overwhelming	1.68	-1.31	-0.23	1.00	2.35	1.93	-1.13	-0.20	0.88	2.09
PROMIS 19	I felt unhappy	2.99	-0.88	0.34	1.54	2.65	3.44	-0.76	0.30	1.37	2.40
PROMIS 20	I felt I had no reason for living	2.34	1.32	2.23	2.74	3.78	2.46	1.20	2.05	2.53	3.54
PROMIS 21	I felt hopeless	3.18	0.31	1.27	2.20	3.01	3.33	0.28	1.16	2.03	2.81
PROMIS 22	I felt ignored by people	1.83	-0.16	0.86	2.02	3.16	2.04	-0.14	0.76	1.81	2.87
PROMIS 23	I felt upset for no reason	2.29	-0.04	0.91	1.91	3.09	2.59	-0.04	0.81	1.71	2.79
PROMIS 24	I felt that nothing was interesting	2.05	0.23	1.32	2.37	3.54	2.22	0.20	1.20	2.17	3.26
PROMIS 25	I felt pessimistic	2.06	-0.56	0.54	1.92	2.90	2.28	-0.49	0.48	1.73	2.64
PROMIS 26	I felt that my life was empty	2.76	0.61	1.37	2.24	2.94	2.96	0.54	1.24	2.05	2.72
PROMIS 27	I felt guilty	1.42	-0.03	1.28	2.69	4.05	1.54	-0.02	1.17	2.46	3.73
PROMIS 28	I felt emotionally exhausted	2.12	-0.44	0.40	1.53	2.39	2.43	-0.38	0.35	1.36	2.13

Item	Equating					Concurrent Calibration				
	a	b1	b2	b3	b4	a	b1	b2	b3	b4
CESD 1	1.36	-0.01	1.23	2.67		1.22	0.19	1.66	3.24	
CESD 2	0.85	0.17	1.74	3.49		0.86	0.32	2.00	3.73	
CESD 3	2.44	0.23	1.05	1.94		2.26	0.36	1.22	2.14	
CESD 4	1.12	0.05	1.60	2.90		1.22	0.15	1.61	2.85	
CESD 5	1.12	-1.01	0.58	2.38		1.06	-1.10	0.59	2.53	
CESD 6	2.74	-0.18	0.72	1.72		2.40	0.08	0.96	1.98	
CESD 7	0.70	-0.90	1.12	3.25		0.73	-0.91	1.16	3.41	
CESD 8	0.89	-0.71	1.41	2.77		1.01	-0.58	1.31	2.74	
CESD 9	1.70	1.08	1.89	2.76		1.78	1.19	2.01	2.88	
CESD 10	1.20	0.02	1.49	3.02		1.31	0.12	1.56	3.04	
CESD 11	0.95	-0.24	1.17	2.57		0.98	-0.19	1.21	2.63	
CESD 12	1.72	-0.20	1.33	2.37		1.73	-0.02	1.47	2.60	
CESD 13	1.23	-0.16	1.15	2.83		1.13	-0.19	1.48	3.21	
CESD 14	1.76	0.01	1.11	2.35		1.80	0.05	1.21	2.38	
CESD 15	1.24	0.19	2.08	4.03		1.13	0.48	2.41	4.26	
CESD 16	1.65	-0.05	1.32	2.36		1.74	0.10	1.40	2.53	
CESD 17	1.62	0.88	1.65	2.78		1.49	1.00	1.90	3.12	
CESD 18	2.14	-0.21	1.01	2.14		2.06	-0.15	1.17	2.32	
CESD 19	1.62	0.46	1.80	3.16		1.57	0.56	1.94	3.25	
CESD 20	1.28	-0.25	1.29	2.89		1.34	-0.13	1.40	2.93	
BDI 1	2.12	0.93	2.54	3.06		2.08	0.94	2.58	3.11	
BDI 2	1.73	1.32	2.72	3.61		1.70	1.33	2.76	3.67	
BDI 3	2.43	1.45	2.29	3.14		2.39	1.46	2.33	3.19	
BDI 4	1.99	0.98	2.30	3.25		1.96	0.99	2.33	3.30	
BDI 5	1.72	1.48	2.66	3.87		1.70	1.50	2.69	3.92	
BDI 6	1.48	1.47	2.53	3.07		1.46	1.49	2.57	3.12	
BDI 7	2.16	0.91	2.32	3.19		2.14	0.91	2.35	3.24	
BDI 8	1.92	0.93	2.25	3.19		1.90	0.94	2.28	3.23	
BDI 9	2.01	1.35	2.59	3.07		1.97	1.36	2.63	3.12	
BDI 10	1.37	1.45	2.52	2.94		1.35	1.47	2.56	2.98	

Item	Equating					Concurrent Calibration				
	<i>a</i>	<i>b1</i>	<i>b2</i>	<i>b3</i>	<i>b4</i>	<i>a</i>	<i>b1</i>	<i>b2</i>	<i>b3</i>	<i>b4</i>
BDI 11	1.40	0.90	2.50	2.93		1.38	0.91	2.54	2.97	
BDI 12	1.44	1.42	3.12	4.11		1.42	1.44	3.17	4.18	
BDI 13	1.87	1.30	2.36	3.56		1.84	1.32	2.40	3.61	
BDI 14	1.58	1.32	2.30	2.85		1.55	1.34	2.34	2.90	
BDI 15	1.93	1.27	2.51	3.65		1.90	1.29	2.54	3.71	
BDI 16	1.10	1.22	3.16	4.42		1.08	1.23	3.20	4.49	
BDI 17	1.22	0.58	3.02	4.35		1.20	0.59	3.07	4.41	
BDI 18	1.22	1.44	2.89	3.81		1.20	1.46	2.93	3.87	
BDI 19	0.57	3.19	5.56	7.44		0.56	3.23	5.64	7.55	
BDI 20	1.25	1.52	3.30	4.41		1.23	1.54	3.35	4.47	
BDI 21	1.11	2.16	3.24	4.45		1.09	2.19	3.29	4.52	

Note: *a* = Discrimination parameter; *b* = difficulty parameters