International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 21(1): 76–85 (2012)
Published online 31 January 2012 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1348

Measuring depression using item response theory: an examination of three measures of depressive symptomatology

THOMAS M. OLINO,[1] LAN YU,[1] DANIEL N. KLEIN,[2] PAUL ROHDE,[3] JOHN R. SEELEY,[3] PAUL A. PILKONIS[1] & PETER M. LEWINSOHN[3]

1 University of Pittsburgh, Pittsburgh, PA, USA
2 Stony Brook University, Stony Brook, NY, USA
3 Oregon Research Institute, Eugene, OR, USA

Key words
item response theory (IRT), Beck Depression Inventory (BDI), Center for Epidemiologic Studies – Depression (CES-D) scale, Schedule for Affective Disorders and Schizophrenia – Children (K-SADS), depression

Correspondence
Thomas M. Olino, Department of Psychiatry, University of Pittsburgh, 3811 O'Hara St, Pittsburgh, PA, 15217–2593, USA. Telephone (+1) 412-383-5433 Fax (+1) 412 -383-5426
Email: olinotm@upmc.edu

Received 23 April 2010;
revised 3 November 2010;
accepted 18 November 2010

Abstract

Evaluations of assessment instruments using classical test theory typically rely on indices of internal consistency, test–retest reliability, and construct validity. However, the use of models from item response theory (IRT) allows comparison of instruments (and items) in terms of the information they provide and where they provide it along the continuum of severity of the construct being assessed. Such results help to identify the measures most appropriate for specific clinical and research contexts. The present study examined the functioning of the Beck Depression Inventory (BDI), the Center for Epidemiologic Studies – Depression (CES-D) scale, and the nine primary symptoms from the depression module of the Schedule for Affective Disorders and Schizophrenia – Children (K-SADS) using IRT methods. A large sample of adolescents ($n = 1709$) completed the BDI, CES-D scale, and K-SADS. IRT calibration analyses demonstrated that the BDI and CES-D scale performed well in similar ranges of depressive severity (approximately −1 to +3 standard deviations [SDs]), although the BDI provided more information at higher severity levels and the CES-D scale at lower severity levels. The K-SADS depression items, which are dichotomous and focused on clinical disorder, provided the least information that was restricted to the narrowest range (approximately +1 to +3 SDs). This work finds consistency between past rationale for the use of the BDI in clinical samples while using the CES-D scale in epidemiological studies. The results for the K-SADS suggest that interview measures may benefit from increasing the number of items and/or response options to collect more psychometric information.

There are many self-report inventories and interview instruments used to assess depression, in particular, and psychopathology, in general. Selection of one instrument over another is often guided by historical preference, psychometric functioning, cost, respondent time demands, training requirements for use of the measure, or knowledge about the nomological network for a particular measure. In an era of heightened attention to empirically based

assessments (EBAs), the use of data to guide measurement selection is of particular importance (Mash and Hunsley, 2005).

Much of the EBA literature (see *Journal of Clinical Child and Adolescent Psychology*, volume 34, issue 3; *Psychological Assessment*, volume 17, issue 3) has relied on classical test theory (CTT), which focuses on internal consistency, test–retest stability, and concurrent and construct validity (Hunsley and Mash, 2007, 2008). Using these indices, many instruments for measuring depression have adequate psychometric properties. However, CTT methods cannot provide direct guidance regarding the ability of a measure to accurately assess depressive symptomatology at various points on the severity continuum. Knowledge about the range of severity assessed by an instrument is crucial for tailoring measurement to address specific questions and addressing them in specific settings. This goal may be achieved through the use of methods from item response theory (IRT) (Embretson and Reise, 2000).

IRT is the basis of modern psychometric techniques that model levels of a latent trait (e.g. depressive symptomatology) as a function of item characteristics, such as item difficulty and discrimination (Embretson and Reise, 2000). IRT methods provide estimates about the position on the latent trait (i.e. theta level) where each item or inventory provides the most information. For example, an item tapping suicidality would likely provide more information about depression at high theta levels than an item about depressed mood. Parameter estimates in IRT models can be integrated to yield item- and test-information functions, which graphically depict the regions of the latent trait continuum assessed most precisely. Item- and test-information functions in IRT are estimated on the same latent trait scale (standardized to have a mean of zero and standard deviation [SD] of one), yielding information that is comparable across inventories (Reise and Henson, 2003). Thus, results from IRT analyses can be used to directly compare multiple measures on a single, common metric.

There has been an increasing number of IRT investigations of measures assessing depressive symptomatology (Bernstein *et al.*, 2006; Carle *et al.*, 2008; Carmody *et al.*, 2006; Cole *et al.*, 2004; Evans *et al.*, 2004; Orlando *et al.*, 2000; Rush *et al.*, 2006; Rush *et al.*, 2003; Santor *et al.*, 1995; Sharp *et al.*, 2006; Stansbury *et al.*, 2006; Trivedi *et al.*, 2004; Uher *et al.*, 2008) and depression-related constructs (Beevers *et al.*, 2007). These studies have had several different goals. First, some have examined the IRT-based psychometric properties of a single instrument (Carle *et al.*, 2008; Cole *et al.*, 2004; Evans *et al.*, 2004; Orlando *et al.*, 2000; Sharp *et al.*, 2006; Stansbury *et al.*, 2006). These investigations provide information about a specific instrument independent of other measures. Second, some studies have examined functioning of different methods (e.g. self-report versus clinician report) using the same instrument (Bernstein *et al.*, 2006; Rush *et al.*, 2006; Rush *et al.*, 2003; Trivedi *et al.*, 2004). These investigations provide information about whether various methods provide different information about the underlying traits at different severity levels. Third, some studies have simultaneously examined the functioning of multiple instruments (Carmody *et al.*, 2006; Santor *et al.*, 1995; Uher *et al.*, 2008). These last investigations provide information about whether different instruments provide valid and precise information across the same or different trait levels.

Assessing depressive disorder and severity often relies on both interview and self-report methods. Uher *et al.* (2008) conducted the only study that included both interview and self-report measures of depression across different instruments. They examined the Hamilton Depression Rating Scale (HAM-D), Montgomery–Asberg Depression Rating Scale (MADRS), and the Beck Depression Inventory (BDI) using IRT methods. The authors did not find satisfactory unidimensionality for the HAM-D and did not provide further information about the measure. They reported that the BDI and MADRS provided complementary information, with the BDI providing more information at higher severity levels, whereas the MADRS provided more information at lower severity levels. With few exceptions (e.g. Carle *et al.*, 2008; Sharp *et al.*, 2006; Uher *et al.*, 2008) previous research has examined the functioning of multiple measures in college-age or adult samples. It is important, however, to consider how instruments assess symptomatology across earlier periods of development (such as adolescence) when clinically significant depressive disorders emerge. Incidence of depressive disorders increases through adolescence (Costello *et al.*, 2002; Hankin *et al.*, 1998), and the lifetime rate of depressive disorder is approximately 25% through age 18 (Lewinsohn *et al.*, 1993).

The present study examines the functioning of three measures of depressive symptomatology in the Oregon Adolescent Depression Project (OADP; Lewinsohn *et al.*, 1993). We focus on two self-report measures of depressive symptoms, the BDI (Beck *et al.*, 1988) and the Center for Epidemiologic Studies – Depression (CES-D) scale (Radloff, 1977). They are two of the most widely used inventories (Myers and Winters, 2002) and their psychometric functioning (using CTT) is highly similar for adolescents and adults (Dougherty *et al.*, 2008). Third, we include the symptom count of the nine primary symptoms from the depression module of the Schedule for

Affective Disorders and Schizophrenia – Children (K-SADS; Orvaschel *et al.*, 1982) as a number of investigators use symptom count as a proxy for depression severity (e.g. Rush *et al.*, 2006). Although this work is largely exploratory, we have some tentative hypotheses. First, given that the K-SADS was designed to elicit information pertaining to clinical diagnosis, we hypothesize that the K-SADS symptom count will best measure depression at the highest severity level. Second, as the K-SADS items are dichotomous and fewer in number, we hypothesize that the K-SADS will provide the least information on depression severity. Third, as the CES-D scale was designed to be used in studies of epidemiology, we hypothesize that the CES-D scale will measure depression over the broadest severity range. Lastly, as the BDI has been useful in assessing changes in depression severity in treatment studies, we hypothesize that the BDI will measure depression severity across higher severity levels than the CES-D scale.

Method

Participants were randomly selected from nine high schools in western Oregon from 1987 to 1989. The participation rate was 61% (Lewinsohn *et al.*, 1993) and comparisons between the 1980 census data and the recruited sample revealed no significant differences on gender, ethnicity, or parental education level. The OADP began with a total of 1709 adolescents (ages 14–18; mean age 16.57, 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) (Lewinsohn *et al.*, 1993). All participants were included in the project, regardless of history of psychopathology.

All participants completed the 20 item CES-D scale (Radloff, 1977) and the 21 item BDI (Beck *et al.*, 1988). The CES-D scale and BDI are two of the most commonly used self-report instruments to assess depression. They have both demonstrated high levels of internal consistency and test–retest stability. Responses for the CES-D scale are frequency based, with four response options that range from rarely/none of the time to most/all of the time. Response options for the BDI are severity based, and include four different severity levels per item. The BDI and CES-D scale assess symptoms as they have been experienced in the past week.

Participants were interviewed with a version of the K-SADS (Orvaschel *et al.*, 1982) that combined features of the Epidemiologic and Present Episode versions. Depressive symptoms were scored as present or absent at a clinically significant level for both the present and worst previous lifetime episode. All depression items were asked of all participants. In addition to being widely used in the literature, the use of symptom counts has been recommended as the primary means of assessing response, remission, recovery, relapse, and recurrence in clinical trials for depression (Rush *et al.*, 2006). Current symptoms of major depressive disorder (MDD), rather than worst past history, were used to maintain comparable time frames for all three measures. The mean interrater reliability (expressed as Kappa) for the individual current major depressive episode (MDE) symptoms was 0.85 ($n = 233$; ranging from 0.79 to 0.90). The point prevalence of MDD was 2.5% ($n = 42$) and the lifetime rate of MDD was 18.4% ($n = 315$).

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 theta (θ) and one or more item parameters (Lord, 1980). A commonly used IRT model for dichotomous items is the two-parameter logistic (2PL) model (Birnbaum, 1968), where two item parameters, a and b, represent item discrimination and item difficulty, respectively. The probability of getting a correct response takes an S-shape, the item characteristic curve (ICC), which provides the probability of choosing an item response category for individuals at different locations on the θ scale. The precise shape of an ICC depends on the values of item parameters. An ICC can be transformed into an item information curve, indicating the amount of psychometric information an item contains at all points along the θ scale. These item information curves can then be added together to form a test information curve, which indicates the amount of information a whole test contains at all points along the θ 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. Another advantage of IRT is that individuals' θ estimates are independent of the number of items or the specific items used for testing.

The most commonly used IRT model for polytomous item responses is the Graded Response Model (GRM) (Samejima, 1969). Analogous to the 2PL model, GRM has one discrimination parameter and a set of difficulty parameters where each parameter is a between-category "threshold." GRM has one discrimination parameter and $n - 1$ threshold parameters for each item, where n is the number of response categories. The discrimination parameter a indicates the shape of the category response curves, with higher discrimination parameters yielding

Int. J. Methods Psychiatr. Res. 21(1): 76–85 (2012). DOI: 10.1002/mpr

steeper curves. Curves that are narrow and peaked indicate that the response categories differentiate well across theta. The *b* difficulty parameters represent item difficulty, which is the theta level at which individuals have a 50% probability of responding affirmatively to the more severe, adjacent response category. For *n* response categories, the GRM estimates $n - 1$ difficulty parameters. For example, if an item has four response options (absent, mild, moderate, and severe), the first *b* parameter represents the theta level at which individuals have a 50% probability of selecting the mild, relative to absent, response category.

Dimensionality

Although IRT has many advantages, IRT models are based on a number of assumptions. One important assumption is that the underlying θ being measured is unidimensional. Prior to examining dimensionality, the sample was randomly divided into two groups of about equal size: a development ($n = 855$) and a validation ($n = 854$) sample. To ensure that the assumption of unidimensionality was met for IRT analysis, we conducted exploratory factor analysis (EFA) of both the items from each measure *and* all items from all measures with the development sample. This was followed by a confirmatory factor analysis (CFA) for each measure and all items from all measures with the validation sample. All factor analyses were run using a mean and variance-adjusted weighted least squares (WLSMV) estimator in Mplus 5.1 (Muthén and Muthén, 1998–2007). Evidence suggests that the WLSMV estimator is optimal for estimating factor analysis parameters with ordinal data (Flora and Curran, 2004). The factor pattern matrix was examined to see how individual items loaded on a single factor.

Following the EFA, the items were then submitted to a single-factor CFA using the validation sample. 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 0.90 or greater for acceptable fit, and 0.95 or greater for good fit. RMSEA values between 0.05 and 0.10 represent an acceptable fit (Steiger, 1990), whereas values less than 0.05 indicate a good fit (McDonald and Ho, 2002). Given these established standards of CFA fit statistics, we also noted the caution of mechanical use of CFA fit criteria as a "permission slip" for modeling data using IRT, since CFA fit results can be affected dramatically by large number of items and skewed data distributions (Cook *et al.*, 2009), which are common characteristics of depression data. Also, we found suggestions in the literature that item parameter estimation is robustness to modest violoations of unidimensionality: Studies using multidimensional data generated by a factor-analytic approach tend to show that a unidimensional IRT model is robust to moderate degrees of multidimensionality (Drasgow and Parsons, 1983; Harrison, 1986; Reckase, 1979).

IRT calibration

We calibrated all items using the GRM for the CES-D scale and BDI and the 2PL model for the K-SADS in Multilog 7.03 (Thissen *et al.*, 2003). The advantage of this simultaneous calibration is that it retains the integrity of the original scales and creates the possibility of adaptive testing using items from all three scales.

Results

In the sample, the mean (and SD) for the BDI, CES-D scale, and K-SADS were 7.05 (7.48), 16.98 (10.62), and 0.76 (1.60), respectively. A small number of participants (0.6% [$n = 10$]) scored a zero on the CES-D scale; a larger number (16.1% [$n = 265$]) scored a zero on the BDI; and a majority (72.0% [$n = 1186$]) scored a zero on the K-SADS.

Assessing dimensionality

For each measure of depressive symptomatology and the pooled items, the scree plot of eigenvalues from the EFA in the development sample was suggestive of a single factor, with the first value substantially larger than the others. Specifically, the ratio of the first to the second eigenvalue was large for the BDI, CES-D scale, K-SADS, and the pooled item set (Table 1). For the BDI, CES-D, and K-SADS measures, all items had factor loadings greater than or equal to 0.30. For the pooled item set, only one item (CES-D item seven: "Everything I did was an effort") had a factor loading (0.28) lower than 0.30.

The one-factor CFA model for the BDI fit well to the validation sample data (Table 1). The one-factor CFA model for the CES-D scale fit adequately to the validation sample data. That is, although the CFI was lower than 0.90, both the TLI and RMSEA were within the acceptable range. The one-factor CFA model for the K-SADS fit well to the validation sample data. The one-factor CFA model for the total item pool fit adequately to the validation sample data. Again, although the CFI was lower than 0.90, both the TLI and RMSEA are within the acceptable range. Overall, these fit indices suggest that the total item

Table 1 Results of unidimensionality analyses for the BDI, CES-D scale, K-SADS, and all items

	EFA		CFA		
	Eigenvalue ratio	Factor determinancy	CFI	TLI	RMSEA
BDI	7.15	0.976	0.965	0.985	0.04
CES-D	5.29	0.967	0.861	0.956	0.09
K-SADS	8.55	0.978	0.995	0.997	0.03
All items	6.60	0.987	0.870	0.942	0.07

pool reflects sufficient unidimensionality for the purposes of calibrating the three measures simultaneously.

IRT analyses

Difficulty and discrimination parameters for all items in the three measures are displayed in Table 2. For the BDI items, discrimination parameters ranged from 0.53 to 2.29 and first, second, and third difficulty parameters ranged from 0.19 to 3.27, 1.84 to 5.44, and 2.38 to 7.37, respectively. For the CES-D items, discrimination parameters ranged from 0.51 to 1.98 and first, second, and third difficulty parameters ranged from −2.38 to 0.93, 0.28 to 2.12, and 1.67 to 4.01, respectively. For the K-SADS items, discrimination parameters ranged from 1.18 to 2.05 and difficulty parameters ranged from 1.53 to 2.97. These item parameters have great impact on item information curves. Curves with higher information along the θ scale, indicated by the height of the curve, have better measurement precision. The location of the peak reflects the discrimination parameter, a measure of which level of severity is best assessed by the item. The range of the difficulty parameters influence how the response categories spread out over the θ range. For example, BDI item one ("Feeling Sad") has a higher discrimination than BDI item six ("Feeling Punished") suggesting that BDI item one provides a higher peak curve than BDI item six, indicating BDI item one providing more information. The range of BDI item one difficulty parameters provides us information that the response categories spread out over 0.78 and 2.77 on the θ scale. In comparison, the range of BDI item six difficulty parameters provides us information that the within response categories spread out over 1.11 to 3.12 on the θ scale. Since the test information function is the sum of the individual item information function, the more information contributed by each individual item, the more information the whole test will provide.

Figure 1 displays the test information function (TIF; Top Panel) and standard error (SE; Bottom Panel) of measurement for the CES-D, BDI, and K-SADS items. The CES-D scale best assessed depressive symptomatology at approximately 1.3 SDs above the mean (as indicated by maximum TIF and minimum SE) and assessed information very well from approximately the mean to 2.5 SDs above the mean. The BDI best assessed depressive symptomatology at approximately 2.5 SDs above the mean (as indicated by maximum TIF and minimum SE) and assessed information very well from approximately the mean to 3.75 SDs above the mean. The K-SADS depression items best assessed depressive symptomatology at approximately 2.25 SDs above the mean (as indicated by maximum TIF and minimum SE); however, these items did not assess information very well at any depression severity level.

Discussion

There is growing recognition of the importance of empirically based assessment for psychopathology (Hunsley and Mash, 2007, 2008). Studies have compared the utility of multiple instruments in a number of ways, however, few studies have compared them using modern measurement techniques, including IRT. In particular, few studies have compared two of the most widely used instruments in clinical and research settings, the CES-D scale and the BDI. Studies have also not examined how information obtained using these self-report measures compares to information obtained from interviews assessing depressive disorder criteria. Lastly, the literature using IRT to examine depressive symptomatology has focused on adult populations. To address these limitations, the present study examined the CES-D scale, BDI, and depressive symptoms from the K-SADS in a large sample of community-residing adolescents using co-calibration IRT methods.

Table 2 IRT calibration of BDI, CES-D, and K-SADS items

		0	1	2	3	Discrimination	Difficulty		
	Descriptor	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*a*	b_1	b_2	b_3
BDI 1	Feel sad	1261 (73.8)	367 (21.5)	45 (2.6)	27 (1.6)	2.18	0.78	2.20	2.77
BDI 2	Discouraged about future	1353 (79.2)	293 (17.1)	45 (2.6)	11 (.6)	1.64	1.14	2.67	3.79
BDI 3	Feel like a failure	1416 (82.9)	215 (12.6)	52 (3.0)	20 (1.2)	2.29	1.17	2.14	2.84
BDI 4	Less satisfaction	1204 (70.5)	414 (24.2)	59 (3.5)	25 (1.5)	1.70	0.73	2.36	3.20
BDI 5	Feel guilty	1341 (78.5)	292 (17.1)	45 (2.6)	25 (1.5)	1.68	1.09	2.48	3.20
BDI 6	Being punished	1292 (75.6)	283 (16.6)	71 (4.2)	57 (3.3)	1.28	1.11	2.38	3.13
BDI 7	Disappointed in myself	1182 (69.2)	442 (25.9)	52 (3.0)	23 (1.3)	2.04	0.62	2.21	2.93
BDI 8	Self-criticism	1098 (64.2)	488 (28.6)	91 (5.3)	25 (1.5)	1.77	0.47	2.07	3.14
BDI 9	Thoughts of killing myself	1285 (75.2)	364 (21.3)	44 (2.6)	8 (0.5)	1.75	0.92	2.63	3.85
BDI 10	Cry more than usual	1298 (76.0)	246 (14.4)	33 (2.1)	121 (7.1)	1.39	1.12	2.15	2.38
BDI 11	More irritated now	1014 (59.3)	481 (28.1)	61 (3.6)	143 (8.4)	1.04	0.49	2.33	2.73
BDI 12	Lost interest in people	1356 (79.3)	283 (16.6)	46 (2.7)	15 (0.9)	1.35	1.29	2.95	4.08
BDI 13	Difficulty making decisions	1291 (75.5)	289 (16.9)	107 (6.3)	14 (0.8)	1.74	0.95	2.05	3.50
BDI 14	Look worse than used to	1329 (77.8)	208 (12.2)	85 (5.0)	76 (4.4)	1.69	1.07	1.84	2.43
BDI 15	Can work about as well	1272 (74.4)	351 (20.5)	73 (4.3)	6 (0.4)	1.60	0.93	2.48	4.22
BDI 16	Can't sleep as well	1126 (65.9)	461 (27.0)	73 (4.3)	40 (2.3)	1.16	0.70	2.69	3.69
BDI 17	Get more tired	951 (55.6)	639 (37.4)	84 (4.9)	25 (1.5)	1.29	0.19	2.52	3.81
BDI 18	Appetite is worse	1261 (73.8)	333 (19.5)	78 (4.6)	29 (1.7)	1.09	1.14	2.86	4.19
BDI 19	Have lost weight	1433 (83.9)	172 (10.1)	61 (3.6)	37 (2.2)	0.53	3.27	5.44	7.37
BDI 20	Worried about my health	1342 (78.5)	294 (17.2)	57 (3.3)	9 (0.5)	1.09	1.42	3.34	5.26
BDI 21	Lost interest in sex	1487 (87.0)	121 (7.1)	70 (4.1)	22 (1.3)	0.92	2.38	3.44	5.09
CESD 1	Bothered by things	777 (45.5)	514 (30.1)	301 (17.6)	113 (6.6)	1.04	−0.25	1.26	2.90
CESD 2	My appetite was poor	809 (47.3)	460 (26.9)	303 (17.7)	133 (7.8)	0.86	−0.18	1.37	3.17
CESD 3	I could not shake off the blues	941 (55.1)	363 (21.2)	256 (15.0)	143 (8.4)	1.86	0.15	0.95	1.86
CESD 4	I am just as good as other people	819 (47.9)	514 (30.1)	252 (14.7)	120 (7.0)	1.21	−0.10	1.34	2.58
CESD 5	I had trouble concentrating	392 (22.9)	579 (33.9)	534 (31.2)	198 (11.6)	1.04	−1.43	0.28	2.25
CESD 6	I felt depressed	683 (40.0)	538 (31.5)	315 (18.4)	169 (9.9)	1.98	−0.37	0.70	1.67
CESD 7	Everything I did was an effort	406 (23.8)	581 (34.0)	465 (27.2)	252 (14.7)	0.51	−2.38	0.70	3.61
CESD 8	I felt good about the future	626 (36.6)	604 (35.3)	305 (17.8)	169 (9.9)	0.87	−0.76	1.23	2.83
CESD 9	I thought I was a failure	1254 (73.4)	270 (15.8)	120 (7.0)	58 (3.4)	1.96	0.81	1.65	2.43
CESD 10	I felt fearful	860 (50.3)	490 (28.7)	244 (14.3)	106 (6.2)	1.28	−0.03	1.29	2.56
CESD 11	My sleep was restless	728 (42.6)	444 (26.0)	319 (18.7)	212 (12.4)	0.96	−0.41	0.92	2.32

Int. J. Methods Psychiatr. Res. 21(1): 76–85 (2012). DOI: 10.1002/mpr

Table 2 *Continued*

		0	1	2	3	Discrimination	Difficulty		
	Descriptor	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*a*	b_1	b_2	b_3
CESD 12	I was happy	765 (44.8)	609 (35.6)	229 (13.4)	102 (6.0)	1.62	−0.21	1.25	2.29
CESD 13	I talked less than usual	677 (39.6)	522 (30.5)	353 (20.7)	153 (9.0)	1.00	−0.53	1.00	2.65
CESD 14	I felt lonely	846 (49.5)	505 (29.5)	245 (14.3)	109 (6.4)	1.51	−0.05	1.17	2.30
CESD 15	People were unfriendly	886 (51.8)	587 (34.3)	186 (10.9)	46 (2.7)	1.00	0.05	2.12	4.01
CESD 16	I enjoyed life	839 (49.1)	516 (30.2)	248 (14.5)	101 (5.9)	1.90	−0.05	1.09	2.11
CESD 17	I had crying spells	1248 (73.0)	251 (14.7)	145 (8.5)	59 (3.5)	1.38	0.93	1.83	2.94
CESD 18	I felt sad	703 (41.1)	598 (35.0)	273 (16.0)	129 (7.5)	1.83	−0.34	0.92	1.95
CESD 19	I felt that people disliked me	999 (58.5)	445 (26.0)	184 (10.8)	76 (4.4)	1.55	0.29	1.49	2.55
CESD 20	I could not get “going”	666 (39.0)	572 (33.5)	328 (19.2)	138 (8.1)	1.23	−0.51	0.96	2.39
K –SADS	Depressed mood	1501 (87.8)	208 (12.2)			1.94	1.53		
K –SADS	Anhedonia	1612 (94.3)	97 (5.7)			2.05	2.06		
K –SADS	Appetite/weight	1518 (88.8)	184 (10.8)			1.29	2.04		
K –SADS	Sleep	1465 (85.7)	238 (13.9)			1.18	1.88		
K –SADS	Agitation/retardation	1598 (93.5)	103 (6.0)			1.80	2.14		
K –SADS	Tired	1529 (89.5)	171 (10.0)			1.25	2.16		
K –SADS	Guilt	1646 (96.3)	56 (3.3)			1.41	2.97		
K –SADS	Concentration	1502 (87.9)	200 (11.7)			1.61	1.71		
K –SADS	Suicidality	1651 (96.6)	51 (3.0)			1.96	2.52		

Int. J. Methods Psychiatr. Res. 21(1): 76–85 (2012). DOI: 10.1002/mpr

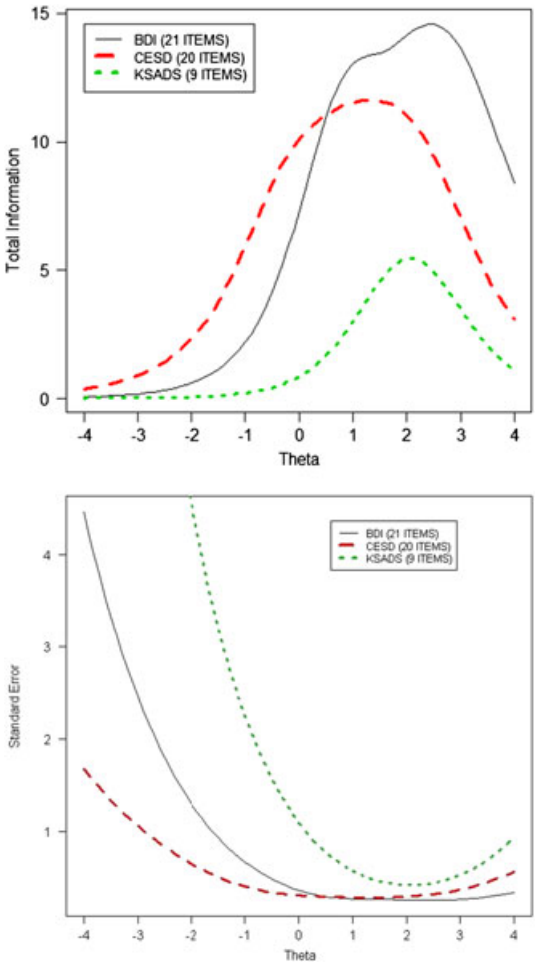

Figure 1 Top Panel: Test information function for CES-D, BDI, and K-SADS depression module items. Bottom Panel: Standard error for CES-D, BDI, and K-SADS depression module items.

We hypothesized that the K-SADS depression items would assess information at the highest depressive severity level because the interview is intended to assess clinically significant symptomatology. The K-SADS items assessed severity at high levels, but not as high as the BDI. We further hypothesized that the K-SADS would assess less total information than the CES-D scale or the BDI because there were fewer items and the items were dichotomous. The K-SADS contributed nine items (compared to 21 contributed by the BDI and 20 contributed by the CES-D scale) and included two response options (compared to four for the BDI and CES-D scale). The results confirmed that the K-SADS depression items provided the least total information. The findings suggest that while diagnostic interviews are essential in making an initial diagnosis of MDD, these types of measures do not provide much information regarding the severity level of the depressive symptomatology. Expanding the response options within the K-SADS items (e.g. by including sub-threshold or severe categories) or using an interviewer-based continuous measure of depressive severity such as the Children's Depression Rating Scale (Poznanski and Mokros, 1996) might provide increased total information. Indeed, Rush *et al.* (2006) found that interview and self-report modalities of the same instrument with multiple response options provided similar amounts of information. Alternatively, one may continue to use the K-SADS to assess and make a diagnosis of a depressive disorder, but assess severity using a separate measure that performs better in that capacity.

We also expected the CES-D scale to assess depressive severity over the broadest depression range compared to the K-SADS and BDI because the CES-D scale was developed to measure depression in epidemiological studies. However, we found that the BDI assessed depression over a similar, but slightly larger range of depression symptom severity than the CES-D scale. Lastly, we expected and found that the BDI provided information at higher depression severity levels than the CES-D scale. Taken together, these results suggest that the CES-D scale and BDI are useful for large ranges of the population. However, the CES-D scale and BDI differed with respect to the level of depressive symptomatology that they optimally assessed. The BDI was more useful for assessing depressive at a higher severity range than the CES-D scale. This suggests that the BDI may be more useful for measuring depressive severity in clinical populations and measuring depressive severity as an index of treatment response. However, the CES-D scale may be more useful for measuring depressive severity in larger, epidemiological samples where the expected average level of depressive is lower. Additionally, the results of the present study may be useful for identifying items that parsimoniously assess depressive symptoms across a wider range of severity than any of the individual scales.

This work benefitted from a large community sample that included individuals with a range of depressive symptomatology assessed using multiple measures. However, the findings should be considered in light of several limitations. First, despite the large sample size, there were a number of response options that were endorsed with low frequency (i.e. less than 5%). Thus, some parameter estimates may be imprecise. Second, tests of unidimensionality were only adequate for the CES-D scale (and for the total item pool, largely due to the influence of the CES-D scale). Additional attention to the

dimensionality of the CES-D scale in adolescent samples may be warranted. However, some work suggests that modest departure from the assumption of unidimensionality does not strongly influence the IRT model and parameter estimates (Dorans and Kingston, 1985; Drasgow and Parsons, 1983). Third, we examined only three of many potential instruments, and the two self-report measures were not specifically designed for use with adolescents, although previous research has found them to be psychometrically valid using CTT indices (Roberts *et al.*, 1991). However, the present results can be used as a bridge to include additional measures of depression to investigate how other measures developed for use with adolescents (e.g. Reynolds, 1987) compare using IRT methods. Fourth, the instruments selected included both self-report and interview based methods. The conclusions may appear to suggest that interview methods (i.e. K-SADS) provide less information than self-reports, however, this issue is confounded in the present study with the number of response options and the nature of the study sample. However, Uher *et al.* (2008) found that the BDI and MADRS provided overlapping information, suggesting that interview methods with multiple response options may provide similar information to self-report measures with multiple response options. Future investigations of comparisons between self-report and interview measures with multiple response options (e.g. the Children's Depression Rating Scale-Revised; Poznanski and Mokros, 1996) will be of particular importance to resolve this issue with respect to adolescents.

In summary, the present study used IRT methods to directly compare the CES-D scale, BDI, and K-SADS depression module on the information provided by each instrument. Largely consistent with expectations, the results suggested differential utility of these instruments and may serve to guide choices about preferred instruments in future research and clinical contexts.

Acknowledgements

This work was supported by the National Institute of Mental Health research award R01 MH40501 to Peter M. Lewinsohn. Support for the first author came from T32 MH 018951.

Declaration of interest statement

The authors have no competing interests.

References

Beck A.T., Steer R.A., Carbin M.G. (1988) Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, **8**, 77–100, DOI: 10.1016/0272-7358(88)90050-5

Beevers C.G., Strong D.R., Meyer B., Pilkonis P.A., Miller I.W. (2007) Efficiently assessing negative cognition in depression: An item response theory analysis of the Dysfunctional Attitude Scale. *Psychological Assessment*, **19**, 199–209, DOI: 10.1037/1040-3590.19.2.199

Bernstein I.H., Rush A., Carmody T.J., Woo A., Trivedi M.H. (2006) Item response analysis of the inventory of depressive symptomatology. *Neuropsychiatric Disease and Treatment*, **2**, 557–564, DOI: 10.2147/nedt.2006.2.4.557

Birnbaum A. (1968) Some latent trait models and their use in inferring an examinee's ability. In Lord F. M., Novick M.R. (eds) Statistical Theories of Mental Test Scores, Reading, MA, Addison-Wessley.

Carle A.C., Millsap R.E., Cole D.A. (2008) Measurement bias across gender on the Children's Depression Inventory: Evidence for invariance from two latent variable models. *Educational and Psychological Measurement*, **68**, 281–303, DOI: 10.1177/0013164407308471

Carmody T.J., Rush A., Bernstein I., Warden D., Brannan S., Burnham D., Woo A., Trivedi M.H. (2006) The Montgomery Asberg and the Hamilton ratings of depression: A comparison of measures. *European Neuropsychopharmacology*, **16**, 601–611, DOI: 10.1016/j.euroneuro.2006.04.008

Cole J.C., Rabin A.S., Smith T.L., Kaufman A.S. (2004) Development and validation of a Rasch-derived CES-D Short Form. *Psychological Assessment*, **16**, 360–372, DOI: 10.1037/1040-3590.16.4.360

Cook K.F., Kallen M.A., Amtmann D. (2009) Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*, **18**, 447–460.

Costello E., Pine D.S., Hammen C., March J.S., Plotsky P.M., Weissman M.M., Biederman J., Goldsmith H., Kaufman J., Lewinsohn P.M., Hellander M., Hoagwood K., Koretz D.S., Nelson C.A., Leckman J.F. (2002) Development and natural history of mood disorders. *Biological Psychiatry*, **52**, 529–542, DOI: 10.1016/S0006-3223(02)01372-0

Dorans N.J., Kingston N.M. (1985) The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, **22**, 249–262.

Dougherty L.R., Klein D.N., Olino T.M., Laptook R.S. (2008) Depression in children and adolescents. In Hunsley J., Mash E.J. (eds) *A Guide to Assessments that Work*, pp. 69–95, New York, Oxford University Press.

Drasgow F., Parsons C.K. (1983) Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, **7**: 189–199, DOI: 10.1177/014662168300700207

Embretson S.E., Reise S.P. (2000) Item Response Theory for Psychologists, Mahwah, NJ, Lawrence Erlbaum Associates.

Evans K.R., Sills T., DeBrota D.J., Gelwicks S., Engelhardt N., Santor D. (2004) An item response analysis of the Hamilton Depression Rating Scale using shared data from two pharmaceutical companies. *Journal of Psychiatric Research*, **38**, 275–284, DOI: 10.1016/j.jpsychires.2003.11.003

Flora D.B., Curran P.J. (2004) An empirical evaluation of alternative methods of estimation for

confirmatory factor analysis with ordinal data. *Psychological Methods*, **9**, 466–491, DOI: 10.1037/1082-989X.9.4.466.

Hankin B.L., Abramson L.Y., Moffitt T.E., Silva P. A., McGee R., Angell K.E. (1998) Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology*, **107**, 128–140.

Harrison D.A. (1986) Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational and Behavioral Statistics*, **11**, 91–115.

Hunsley J., Mash E.J. (2007) Evidence-based assessment. *Annual Review of Clinical Psychology*, **3**, 29–51, DOI: 10.1146/annurev.clinpsy.3.022806.091419

Hunsley J., Mash E.J. (2008) A Guide to Assessments that Work, New York, Oxford University Press.

Lewinsohn P.M., Hops H., Roberts R.E., Seeley J. R., Andrews J.A. (1993) Adolescent psychopathology: I. Prevalence and incidence of depression and other DSM-III-R disorders in high school students. *Journal of Abnormal Psychology*, **102**, 133–144.

Lord F.M. (1980) Applications of Item Response Theory to Practical Testing Problems, Hillsdale, NJ, Lawrence Erlbaum.

Mash E.J., Hunsley J. (2005) Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child & Adolescent Psychology*, **34**, 362–379, DOI: 10.1207/s15374424jccp3403_1

McDonald R.P., Ho M.H.R. (2002) Principles and practice in reporting structural equation analyses. *Psychological Methods*, **7**, 64–82, DOI: 10.1037/1082-989X.7.1.64

Muthén L.K., Muthén B.O. (1998–2007) Mplus User's Guide, 5th edition, Los Angeles, CA, Muthén & Muthén.

Myers K., Winters N.C. (2002) Ten-year review of rating scales. *II. Scales for internalizing disorders. Journal of the American Academy of Child and Adolescent Psychiatry*, **41**, 634–659, DOI: 10.1097/00004583-200206000-00004

Orlando M., Sherbourne C.D., Thissen D. (2000) Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, **12**, 354–359, DOI: 10.1037//1040-3590.12.3.354

Orvaschel H., Puig-Antich J., Chambers W.J., Tabrizi M.A., Johnson R. (1982) Retrospective assessment of prepubertal major depression with the Kiddie-SADS-E. *Journal of the American Academy of Child and Adolescent Psychiatry*, **21**, 392–397.

Poznanski E., Mokros H. (1996) Children's Depression Rating Scale – Revised Manual, Los Angeles, CA, Western Psychological Services.

Radloff L.S. (1977) The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, **1**, 385–401.

Reckase M.D. (1979) Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, **4**, 207–230.

Reise S.P., Henson J.M. (2003) A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, **81**, 93–103, DOI: 10.1207/S15327752JPA8102_01

Reynolds W. (1987) Professional Manual for the Reynolds Adolescent Depression Scale, Odessa, FL, Psychological Assessment Resources, Inc.

Roberts R.E., Lewinsohn P.M., Seeley J.R. (1991) Screening for adolescent depression: A comparison of depression scales. *Journal of American Academy of Child and Adolescent Psychiatry*, **30**, 58–66.

Rush A.J., Bernstein I.H., Trivedi M.H., Carmody T.J., Wisniewski S., Mundt J.C., Shores-Wilson K., Biggs M.M., Woo A., Nierenberg A.A., Fava M. (2006) An evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: A sequenced treatment alternatives to relieve depression trial report. *Biological Psychiatry*, **59**, 493–501, DOI: 10.1016/j.biopsych.2005.08.022

Rush A.J., Trivedi M.H., Ibrahim H.M., Carmody T.J., Arnow B., Klein D.N., Markowitz J.C., Ninan P. T., Kornstein S., Manber R., Thase M.E., Kocsis J. H., Keller M.B. (2003) 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*, **54**, 573–583, DOI: 10.1016/S0006-3223(02)01866-8

Samejima F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, **17**, 1–100.

Santor D.A., Zuroff D.C., Ramsay J.O., Cervantes P., Palacios J. (1995) Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment*, **7**, 131–139.

Sharp C., Goodyer I.M., Croudace T.J. (2006) The Short Mood and Feelings Questionnaire (SMFQ): A unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *Journal of Abnormal Child Psychology*, **34**, 379–391, DOI: 10.1007/s10802-006-9027-x

Stansbury J.P., Ried L., Velozo C.A. (2006) Unidimensionality and bandwidth in the Center for Epidemiologic Studies-Depression (CES-D) scale. *Journal of Personality Assessment*, **86**, 10–22, DOI: 10.1207/s15327752jpa8601_03

Steiger J.H. (1990) Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, **25**, 173–180.

Thissen D., Chen W.-H., Bock R.D. (2003) MULTILOG 7 for Windows: Multiple-Category Item Analysis and Test Scoring Using Item Response Theory [Computer Software]. Lincolnwood, IL, Scientific Software International, Inc.

Trivedi M.H., Rush A.J., Ibrahim H.M., Carmody T.J., Biggs M.M., Suppes T., Crismon M.L., Shores-Wilson K., Toprac M.G., Dennehy E. B., Witte B., Kashner T.M. (2004) The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: A psychometric evaluation. *Psychological Medicine*, **34**, 73–82, DOI: 10.1017/S0033291703001107

Uher R., Farmer A., Maier W., Rietschel M., Hauser J., Marusic A., Mors O., Elkin A., Williamson R.J., Schmael C. (2008) Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychological Medicine*, **38**, 289–300, DOI: 10.1017/S0033291707001730