Psychological Assessment

Structure and Measurement of Depression in Youths: Applying Item Response Theory to Clinical Data

David A. Cole, Li Cai, Nina C. Martin, Robert L. Findling, Eric A. Youngstrom, Judy Garber, John F. Curry, Janet S. Hyde, Marilyn J. Essex, Bruce E. Compas, Ian M. Goodyer, Paul Rohde, Kevin D. Stark, Marcia J. Slattery, and Rex Forehand

Online First Publication, May 2, 2011. doi: 10.1037/a0023518

Psychological Assessment

© 2011 American Psychological Association
1040-3590/11/$12.00 DOI: 10.1037/a0023518

Structure and Measurement of Depression in Youths: Applying Item Response Theory to Clinical Data

David A. Cole
Vanderbilt University

Li Cai
University of California, Los Angeles

Nina C. Martin
Vanderbilt University

Robert L. Findling
Case Western Reserve University

Eric A. Youngstrom
University of North Carolina at Chapel Hill

Judy Garber
Vanderbilt University

John F. Curry
Duke University School of Medicine

Janet S. Hyde
University of Wisconsin–Madison

Marilyn J. Essex
University of Wisconsin School of Medicine and Public Health

Bruce E. Compas
Vanderbilt University

Ian M. Goodyer
University of Cambridge

Paul Rohde
Oregon Research Institute

Kevin D. Stark
University of Texas at Austin

Marcia J. Slattery
University of Wisconsin School of Medicine and Public Health

Rex Forehand
University of Vermont

Our goals in this article were to use item response theory (IRT) to assess the relation of depressive symptoms to the underlying dimension of depression and to demonstrate how IRT-based measurement strategies can yield more reliable data about depression severity than conventional symptom counts. Participants were 3,403 children and adolescents from 12 contributing clinical and nonclinical samples; all participants had received the Kiddie Schedule of Affective Disorders and Schizophrenia for School-Aged Children. Results revealed that some symptoms reflected higher levels of depression and were more discriminating than others. Furthermore, use of IRT-based information about symptom severity and discriminability in the measurement of depression severity was shown to reduce measurement error and increase measurement fidelity.

Keywords: item response theory (IRT), depression, children, adolescents, Kiddie–SADS (K–SADS)

David A. Cole, Nina C. Martin, Judy Garber, and Bruce E. Compas, Department of Psychology and Human Development, Vanderbilt University; Li Cai, Departments of Education and Psychology, University of California, Los Angeles; Robert L. Findling, Department of Psychiatry, Case Western Reserve University; Eric A. Youngstrom, Departments of Psychology and Psychiatry, University of North Carolina at Chapel Hill; John F. Curry, Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine; Janet S. Hyde, Department of Psychology, University of Wisconsin–Madison; Marilyn J. Essex and Marcia J. Slattery, Department of Psychiatry, University of Wisconsin School of Medicine and Public Health; Ian M. Goodyer, Department of Psychiatry, University of Cambridge, Cambridge, England; Paul Rohde, Oregon Research Institute, Eugene, Oregon; Kevin D. Stark, Department of Educational Psychology, University of Texas at Austin; Rex Forehand, Department of Psychology, University of Vermont.

continued

The application of item response theory (IRT) to semistructured clinical interview data can simultaneously advance the understanding of psychopathology and enhance the fidelity of its measurement. IRT has proven useful when applied to paper-and-pencil measures of depressive symptoms (Bedi, Maraun, & Chrisjohn, 2001; Cassano et al., 2009; Sharp, Goodyer, & Croudace, 2006; Waller, Compas, Hollon, & Beckjord, 2005). For clinical researchers, the closest thing to a gold standard for the assessment of child and adolescent depression is a semistructured clinical interview, typically administered not just to the child but to a parent or other caregiver as well. As such, the semistructured clinical interview is inherently a multimethod assessment system, filtering information from multiple informants through interviewers with clinical training and expertise. Analyzing symptom-level information derived from such measures can provide insights into the structure of the underlying depression construct, lead to the psychometric enhancement of these measures, and eventually enable researchers to derive more information from such interviews of depressed children and adolescents. Although IRT analyses have been conducted with adult samples (e.g., Simon & Von Korff, 2006), relatively few IRT analyses of clinical interview data have been conducted with child or adolescent populations (e.g., Small et al., 2008).

In both child and adult populations, conventional factor analyses have informed researchers' understanding about the relation of specific depression symptoms to the underlying latent variable (Aggen, Neale, & Kendler, 2005; Ryan et al., 1987). IRT provides at least three additional kinds of information. First, in IRT, each symptom is linked to a specific level of depression severity. Consider an analogy. On a math test, some items may be more difficult than others, such that passing a more difficult item may suggest that the respondent has a higher level of math ability than does passing an easier item. The same may be true for depressive symptoms. Some symptoms may be evident at relatively mild levels of the disorder, whereas other symptoms may only emerge at very severe levels. In other words, severe depression may be characterized by symptoms that are not often evident in mild depression. If the severity of depression is assessed simply by counting the number of symptoms, then all symptoms are treated as though they were of equal severity or importance and other valuable information that could be derived from the assessment process potentially is ignored.

Second, IRT allows for the possibility that all symptom ratings may not be equally reliable or discriminating indicators of depression. Some symptoms may be strong indicators of depression, constituting core characteristics of the disorder. Other symptoms may be less strongly related to a depressive disorder or may be relatively nonspecific signs of the disorder. Unlike methods based on classical test theory, IRT-based estimates of item (or symptom) discriminability are not sample dependent once the IRT model is calibrated (Reise & Waller, 2009). That is, the psychometric properties of the items do not vary from sample to sample but generalize to all samples from the same population, revealing something about the structure of the underlying latent dimensions in general. Furthermore, utilization of IRT-derived discriminability information (in conjunction with severity information) can greatly enhance the fidelity of the information that can be derived from clinical interview data.

Third, the application of IRT to a collection of symptoms enables researchers to ascertain the degree to which a measure "covers" the latent variable. That is, IRT reveals how informative a measure is at all levels of the underlying dimension. Some measures may be particularly discriminating at the high end of depression severity and be especially useful in clinical settings. Other measures may be maximally discriminating at the low end of depression severity and be useful as a screening device in nonclinical populations. A measure used in clinical trials should be discriminating along the entire range of severity, because participants typically start at very high levels of the disorder but (it is hoped) end up at much lower levels.

When symptoms of a disorder (as assessed by a semistructured clinical interview) are treated as "items" in an IRT analysis, these three kinds of information simultaneously serve two purposes. First, they provide more information about the relations of symptoms to the underlying depression factor(s). And second, they can

This research was supported in part by the following grants. Li Cai: Institute of Education Sciences Grants R305B080016 and R305D100039 and National Institute on Drug Abuse Grants R01DA026943 and R01DA030466; David Cole: National Institute of Mental Health (NIMH) Grant R01MH64650 and by gifts from Patricia and Rodes Hart and from the Warren family; Bruce Compas: NIMH Grants R01MH069940 and R01MH069928 and a gift from Patricia and Rodes Hart; Robert Findling and Eric Youngstrom: NIMH Grants R01MH066647 and P20-MH066054 and a Clinical Research Center Grant from the Stanley Medical Research Institute; Rex Forehand: NIMH Grants RO1MH069940 and RO1MH069928 and a gift from the Heinz and Rowena Ansbacher Professorship; Janet S. Hyde and Marilyn J. Essex: John D. and Catherine T. MacArthur Foundation Research Network on Psychopathology and Development and NIMH Grants R01MH44340, P50-MH52354, P50-MH69315, and P50-MH84051; Judy Garber: NIMH Grants R29MH454580 and R01MH57822 and National Institute of Child Health and Human Development Grant P30-HD15052, Faculty Scholar Award 1214-88 and Grant 173096 from the William T. Grant Foundation, and Independent Scientist Award K02MH66249; Ian Goodyer: National Health Service (NHS) Health Technology Assessment Programme, Central Manchester and Manchester Children's University Hospitals NHS Trust, and the Cambridge and Peterborough Mental Health Trust; John S. March: NIMH 98-DS-0008 (Treatment for Adolescents With Depression Study [TADS]; John F. Curry was a co-investigator who collaborated on this project.); Paul Rohde: NIMH Grants MH56238, MH67183, and MH 56238; Marcia J. Slattery: Grant 1UL1RR025011 from the Clinical and Translational Science Award Program of the National Center for Research Resources in NIH and NIMH grant P50-MH69315; Kevin Stark: NIMH Grant R01MH63998; Myrna Weissman: NIMH Grant R01MH063852 and NIMH Contract N01 MH90003.

Robert Findling receives or has received research support, acted as a consultant, and/or served on a speaker's bureau for Abbott, Addrenex, AstraZeneca, Biovail, Bristol-Myers Squibb, Forest, GlaxoSmithKline, Johnson & Johnson, KemPharm Lilly, Lundbeck, Neuropharm, Novartis, Noven, Organon, Otsuka, Pfizer, Rhodes Pharmaceuticals, Sanofi-Aventis, Schering-Plough, Seaside Therapeutics, Sepracore, Shire, Solvay, Sunovion, Supernus Pharmaceuticals, Validus, and Wyeth.

Correspondence concerning this article should be addressed to David A. Cole, Department of Psychology and Human Development, Vanderbilt University, Peabody MSC 512, 230 Appleton Place, Nashville, TN 37203. E-mail: david.cole@vanderbilt.edu

be used in the construction of new indices (and even computer adaptive testing methods) that are more efficient and more discriminating across a wider range of the targeted dimension. IRT has often resulted in tests that are shorter and more sensitive to the detection of individual differences (Gibbons et al., 2008; Reeve, Burke, et al., 2007; Reeve, Hays, et al., 2007). Clinical applications of IRT are rare, largely because IRT requires sample sizes that are substantially larger than are available in most clinical data bases. Of the few such studies that do exist, almost all have focused on paper-and-pencil measures of psychopathology, on which large samples can more easily be obtained (Cassano et al., 2009; Fliege et al., 2009; Gardner et al., 2004; Gibbons et al., 2008). To solve the sample size problem, we aggregated data from clinical researchers in the United States and Great Britain who used the Kiddie Schedule of Affective Disorders and Schizophrenia for School-Aged Children (K–SADS) to measure major depressive disorder (MDD) in children and adolescents. We intentionally sought a wide variety of data sets including community samples, high-risk samples, and clinical treatment samples, so that collectively they would represent all levels of depression severity. We also sought samples that would contribute to the demographic diversity of the composite data set, in terms of age, sex, and ethnicity.

Thus, in the present study, we had three goals or hypotheses. First, we anticipated that the presence of some depressive symptoms would reflect a more severe underlying depressive disorder than would the presence of other symptoms. For example, we hypothesized that depressed mood would be a relatively mild symptom (as it is widely regarded as the core or gateway symptom of MDD), whereas suicidal ideation would be a more severe symptom, tending to manifest itself at relatively severe levels of the disorder. Second, we expected items to evince different levels of discriminability, with some being highly reflective of the underlying disorder (e.g., anhedonia; Clark & Watson, 1991; Lonigan, Carey, & Finch, 1994) and others being only moderately reflective of the condition (e.g., weight or appetite disturbance)—perhaps because they are also characteristic of other disorders. Finally, we sought to examine the degree to which an IRT-based scoring of the K–SADS would yield more reliable symptom ratings and would generate more information than conventional methods of scoring the K–SADS to measure depression severity.

Method

Data Set Selection

Three criteria were required for a data set to be included in the study. First, it had to contain symptom-level information either about participants' current state or their recent episode of MDD, derived from K–SADS interviews with children and parents. Second, participants had to be from 5 to 18 years old. Third, the K–SADS data must have been collected prior to any treatment or preventive intervention. Prior to data acquisition, we obtained institutional review board approval, arranged for the complete de-identification of data sets, made explicit the limitations on our use of the data, conferred with the principal investigator (PI) and other study collaborators to ensure that no conflicts of interest existed between our research agenda and those of the original investigator(s), discussed authorship, and obtained signed letters of agreement from the PI or co-PI of each project.

In total, we obtained 12 different data sets, yielding a total of 3,403 participants. We refer to each study by the investigator who was our key collaborator on this project. When this person provided access to multiple data sets, we indicate the study title as well. Contributors included the following: Cole (Cole et al., in press), Compas and Forehand (Compas et al., 2009; 2010), Curry (Treatment for Adolescents With Depression Study [TADS], 2003, 2005), Findling (Findling et al., 2005), Garber (multiple data sets: Garber 1 indicates the Development of Depression Project [DODP], Gallerani, Garber, & Martin, 2010; Garber & Cole, 2010; and Garber, Keiley, & Martin, 2002; Garber 2 indicates Parent–Child Project [PCP], Garber, Ciesla, McCauley, Diamond, & Schloredt, 2011), Goodyer (Goodyer et al., 2007, 2008), Hyde and Essex (Essex et al., 2006; 2009; Grabe, Hyde, & Lindberg, 2007; Mezulis, Priess, & Hyde, 2010; Priess, Lindberg, & Hyde, 2009), Rohde (N. Kaufman, Rohde, Seeley, Clarke, & Stice, 2005; Rohde, Clarke, Mace, Jorgensen, & Seeley, 2004; Rohde, Seeley, Kaufman, Clarke, & Stice, 2006), Stark (Fisher, 2010), Weissman (Pilowsky et al., 2008; Weissman, Pilowsky, & Wickramaratne, 2006), and Youngstrom (Youngstrom et al., 2005). Key characteristics of the data sets appear in Table 1.

Measures

Several versions of the K–SADS were used in the contributing studies. These included K–SADS–Present and Lifetime Version (K–SADS–PL; J. Kaufman et al., 1997), K–SADS–PL Version 1.0 (J. Kaufman, Birmaher, Brent, Rao, & Ryan, 1996), K–SADS–Epidemiological Version (K–SADS–E; Orvaschel, 1994), Washington University in St. Louis K–SADS (WASH–U–K–SADS; Geller, Zimerman, & Williams, 2001), and K–SADS–Version IV–Revised (K–SADS–IV–R; Ambrosini & Dixon, 1996). When K–SADS data were available for multiple episodes of major depression, we focused on the current or most recent episode. All five K–SADS versions provide lines of inquiry and example questions for interviewers to use with children (about their own symptoms) and with parents (about their child's symptoms). Slight differences exist in the example questions; however, no version requires that the interviewer adhere to the exact questions that are listed. In fact, all versions recommend that interviewers utilize their clinical skills to probe in ways that the participants can understand.

Because the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; American Psychiatric Association, 2000) regards irritability or anger as evidence of mood disturbance in children, we treated it as a separate symptom in the current study. We pooled questions for assessment of the depressive symptoms across the five versions of the K–SADS; examples of these questions include:

1. Depressed mood. Have you ever felt sad, blue, down, or empty? Did you feel like crying? Did you have a bad feeling all the time that you couldn't get rid of?

2. *Irritability or anger*. Was there ever a time when you got annoyed, irritated, or cranky at little things? Did you ever have a time when you lost your temper a lot?

Table 1
Sample Characteristics by Study

			Age (years)									
Study	*N*	Sample	<8	8–9	10–11	12	13	14	15	16	>16	Missing
Cole	100	High-risk: Cognitive vulnerability for depression	2	44	37	13	3					1
Compas & Forehand	242	High-risk: Children of depressed parents		55	67	37	36	22	24	1		
Curry	439	Targeted: Adolescents with major depressive disorder				48	73	81	91	93	53	
Findling	851	Targeted: Clinic referred	121	170	164	71	60	60	63	55	69	18
Garber 1	240	High-risk: Children of depressed mothers			56	163	20	1				
Garber 2	227	High-risk: Children of depressed parents	7	28	48	36	43	30	19	14	2	
Goodyer	208	Targeted: Clinic referred for depression			5	12	32	65	65	25	4	
Hyde & Essex	245	Community: No risk factors						26	125	90	1	3
Rohde	182	Targeted: Comorbid for MDD and conduct disorder				6	15	47	41	37	36	
Stark	98	Targeted: Girls with depressive disorders		18	47	24	8	1				
Weissman	151	High-risk: Children of depressed mothers	11	32	32	17	20	12	15	8	4	
Youngstrom	420	Targeted: Clinic referred	98	67	70	47	41	24	28	23	17	5
Totals	3,403		239	414	526	474	351	369	471	346	186	27

Note. **Cole** = Cole et al., in press; **Compas & Forehand** = Compas et al., 2009, 2010; **Curry** = Treatment for Adolescents With Depression Study [TADS], 2003, 2005; **Findling** = Findling et al., 2005; **Garber 1** = Development of Depression Project, Gallerani, Garber, & Martin, 2010; Garber & Cole, 2010; Garber, Keiley, & Martin, 2002; **Garber 2** = Parent–Child Project, Garber, Ciesla, McCauley, Diamond, & Schloredt, 2011); **Goodyer** = Goodyer et al., 2007, 2008; **Hyde & Essex** (combined data set) = Essex et al., 2006, 2009; Grabe, Hyde, & Lindberg, 2007; Mezulis, Priess, & Hyde, 2010; Priess, Lindberg, & Hyde, 2009); **Rohde** = N. Kaufman, Rohde, Seeley, Clarke, & Stice, 2005; Rohde, Clarke, Mace, Jorgensen, & Seeley, 2004; Rohde, Seeley, Kaufman, Clarke, & Stice, 2006; **Stark** = Fisher, 2010; **Weissman** = Pilowsky et al., 2008; Weissman, Pilowsky, & Wickramaratne, 2006; **Youngstrom** = Youngstrom et al., 2005.

3. *Pervasive anhedonia* (lack of interest, apathy, low motivation, or boredom). Has there ever been a time you felt bored a lot of the time? Did you have to push yourself to do your favorite activities? Did they interest you?

4. *Weight or appetite disturbance*. (a) Appetite loss: How is your appetite? Do you feel hungry often? Do you leave food on your plate? Do you sometimes have to force yourself to eat? (b) Weight loss: Have you lost any weight since you started feeling sad? Do you find your clothes are looser now? (c) Appetite gain: Have you been eating more than before? Is it like you feel hungry all the time? (d) Weight gain: Have you gained any weight since you started feeling sad? Have you had to buy new clothes because the old ones did not fit any longer?

5. *Sleep disturbance*. (a) Insomnia: Do you have trouble sleeping? How long does it take you to fall asleep? Do you wake up in the middle of the night? Do you wake up earlier than you have to? (b) Hypersomnia: Are you sleeping longer than usual? Do you go back to sleep after you wake up in the morning?

6. *Psychomotor disturbance*. (a) Agitation: Since you've felt sad, are there times when you can't sit still, or you have to keep moving and can't stop? Do people tell you not to talk so much? (b) Retardation: Since you started feeling sad, have you noticed that you can't move as fast as before? Has your speech slowed down? Have you felt like you are moving in slow motion?

7. *Fatigue, lack of energy, or tiredness*. Have you been feeling tired? Do you take naps because you feel tired? Do you have to rest? Do your limbs feel heavy? Is it very hard to get going? . . . to move your legs?

8. *Self-perceptions*. (a) Worthlessness: How do you feel about yourself? Do you like yourself? Do you ever think of yourself as pretty or ugly? Do you think you are bright or stupid? (b) Excessive or inappropriate guilt: Do you feel guilty about things you have not done? Or are actually not your fault? Do you feel you cause bad things to happen? Do you think you should be punished for this?

9. *Cognitive disturbance*. (a) Concentration, inattention, slowed thinking: Sometimes children have a lot of trouble concentrating, like [list examples]. Have you been having this kind of trouble? Is your thinking slowed down? When you try to concentrate on something, does your mind drift off to other thoughts? Can you pay attention in school? Can you pay attention when you want to do something you like? (b) Indecision: When you were feeling sad, was it hard for you to make decisions?

10. *Suicide*. Sometimes children who get upset or feel bad wish they were dead or feel they'd be better off dead. Have you ever had these types of thoughts? Sometimes children who get upset or feel bad think about dying or even killing themselves. Have you ever had such thoughts? How would you do it? Did you have a plan?

All versions had good interrater reliability in the studies that contributed data. Previously accumulated validity information supports the use of all versions of the K–SADS to measure and diagnose depression (Ambrosini, 2000). The K–SADS–PL and

K–SADS–E versions are primarily categorical diagnostic interviews, whereas the WASH–U–K–SADS and K-SADS–IV–R measure symptom severity and are sometimes used to measure degree of treatment response (Ambrosini, 2000). The various versions of the K–SADS also differ in the scaling used to quantify symptom severity. The K–SADS–PL has a 3-point scale, where 1 = *symptom is absent*, 2 = *symptom is present at a subclinical level,* and 3 = *symptom is severe and frequent enough to be at or above threshold*. Other versions of the interview have 4-, 6-, and 7-point scales. All versions provide explicit severity and frequency anchors for their scales. These anchors enabled us to translate all measures onto the 3-point K–SADS–PL scale. We converted the 6-point K–SADS–IV–R scale such that 1–2 = 1, 3 = 2, and 4–6 = 3.[1] We converted the 4-point Orvaschel versions of the K–SADS such that 1 = 1, 2 = 2, and 3–4 = 3. We converted a 7-point version of the K–SADS–PL such that 1–3 = 1, 4 = 2, and 5–7 = 3. And we modified the WASH-U–K–SADS such that 1–2 = 1, 3 = 2, and 4–7 = 3. We used a multigroup approach in our data analytic method (which we will discuss later), enabling us to confirm the psychometric equivalence of the resultant scales across studies.

Variables. We extracted four kinds of variables from the K–SADS data. The first was a collection of symptom-specific variables (on the 3-point scales described earlier). The second was a dichotomous index of presence or absence of MDD, reflecting *DSM–IV–TR* criteria (using only "above-threshold" symptoms). Third was a *raw symptom count* variable, ranging from 0–10, reflecting presence or absence of the 10 depression symptoms (also using only above-threshold symptoms). Fourth was a *raw symptom sum* variable, equal to the sum of the 10 symptom-specific (3-point) variables.

Missing data. Three different patterns of missing data occurred across the contributing data sets. Pattern 1 (10% of the cases) emerged because in some studies, questions about depressed mood, irritability, and anhedonia were used as screening questions, and the remaining depressive symptoms were not addressed (presumably because they did not meet criteria on the screening symptoms). Pattern 2 (12.5%) emerged because in some studies, participants were asked the first screening questions plus the suicide screening question but were not asked about other symptoms. Pattern 3 (5%) consisted of random missing data. Comparisons of participants with each pattern of missing data with the larger pool of participants with no missing data revealed no psychometric differences between the groups. Consequently, we did not exclude participants with partial data but used an expectation-maximization (EM) algorithm for the multiple group full-information maximum marginal likelihood estimation that utilized all available data (Bock & Aitkin, 1981).

Results

Descriptive Statistics

Overall, the composite data set contained information on 1,722 boys and 1,678 girls (gender data were missing for three participants). Ages ranged from 5 to 18 years (*M* = 12.39, *SD* = 2.99). See Table 1. The sample was ethnically diverse: with 66% White, 24% African American, 4% Hispanic, and 6% other. All were English speaking. Means and *SD*s for all symptom variables and the total symptom count appear in Table 2.

Table 2
Sample Descriptive Statistics for Variables Derived From the Kiddie Schedule for Affective Disorders and Schizophrenia

Variable	*M*	*SD*
Affective disturbance		
Depressed mood	1.69	0.86
Irritability	1.68	0.85
Anhedonia	1.69	0.89
Weight or appetite disturbance	1.69	0.89
Sleep disturbance	1.93	0.94
Psychomotor disturbance	1.84	0.90
Fatigue, lack of energy, or tiredness	1.89	0.94
Feelings of worthlessness or guilt	1.97	0.92
Cognitive disturbance	1.97	0.94
Suicide	1.49	0.78
Raw symptom count (0–10)	2.66	3.06
Raw symptom sum (10–30)	14.93	8.01
Major depression (0, 1)	0.33	0.47

Note. Analytic sample *N* = 3,403. Variables coded 1–3, unless otherwise specified.

Testing Unidimensionality and Local Independence

Two closely related assumptions of IRT are unidimensionality of the symptoms and the absence of noteworthy local dependencies between the symptoms after accounting for the primary underlying factor (Reise & Waller, 2009). We used categorical weighted least squares confirmatory factor analysis and IRT methods to test these assumptions. Specifically, we constrained all symptoms to load only onto a single underlying factor, allowing no correlations among the disturbances. Although the overall chi-square was significant, $\chi^2(35) = 142.71$, $p < .001$, other fit indices clearly revealed that the fit was excellent: comparative fit index (CFI) = 0.99, normed fit index (NFI) = 0.99, root-mean-square error of approximation (RMSEA) = 0.035, 90% confidence interval (CI) [0.030, −0.059], suggesting that the model fit the data well (Browne & Cudeck, 1993). Factor loadings appear in Table 3. Further, the root-mean square of the residuals was only 0.036. Eigenvalues of the estimated polychoric correlation matrix were 7.54, 0.51, 0.41, 0.30, 0.29, 0.26, 0.23, 0.19, 0.15, and 0.12. Taken together, these results provide strong support for the unidimensionality of the depressive symptoms. We also conducted an exploratory full-information factor analysis (Bock, Gibbons, & Muraki, 1988) using IRT for Patient-Reported Outcomes (IRTPRO; Cai, du Toit, & Thissen, in press) software. Extracting two factors (in an oblique, direct quartimin rotation) revealed evidence of overfactoring (i.e., the second factor had only one large loading, as shown in Table 3). Finally, Chen and Thissen's (1997) local dependence indices showed no discernable pattern across all item pairs, suggesting no evidence of nuisance factors.

IRT Analyses

General analytic approach. Our primary analytic approach consisted of a multigroup, unidimensional, graded IRT model. We arbitrarily selected one of the contributing data sets (Garber 2 = PCP)

[1] Consultation with experts suggested one exception. Suicide was scaled such that 1 = 1, 2 = 2, and 3–6 = 3.

Table 3
Factor Loadings From One- and Two-Factor Analyses of 10 Depression Symptoms

Symptom	One-factor confirmatory solution	Rotated exploratory two-factor solution	
		Factor 1	Factor 2
Depressed mood	0.95	0.83	0.14
Irritability	0.88	0.80	0.09
Anhedonia	0.91	0.97	−0.09
Weight or appetite disturbance	0.82	0.54	0.27
Sleep disturbance	0.87	0.80	0.04
Psychomotor disturbance	0.88	0.73	0.09
Fatigue, lack of energy, or tiredness	0.91	0.98	−0.12
Feelings of worthlessness or guilt	0.84	0.64	0.20
Cognitive disturbance	0.93	0.90	−0.02
Suicide	0.73	0.13	0.72

Note. Standard errors for all confirmatory factor analyses were 0.01.

to serve as the reference group in this analysis. We used Samejima's (1969) graded response model because it is specifically suited to examining the 3-point ratings for each symptom (absent, subclinical, clinical). We used IRTPRO to estimate these models. We relied on Orlando and Thissen's (2000) summed-score item-fit statistics and plots to test the misfit in the shape of item response characteristic curves. In every case, we found that the model-expected probabilities closely followed the observed response probabilities.

Cross-study comparisons. By design, we selected highly heterogeneous data sets. Examining them directly in a multiple-group model, we demonstrated that we can successfully capture this heterogeneity (see Figure 1).[2] Note that all distributions are plotted on a common metric for the latent depression variable. In IRT (as in common factor analysis), this metric is arbitrary. In the current application, we set the reference group mean at 0 and the *SD* at 1. We then mapped all the other groups onto this metric. Because many of the other groups contained more seriously depressed participants, the mean and *SD* of the combined sample were greater than those for the reference group. For the combined sample, the mean of the IRT scale score was 2.60, and the *SD* was 1.28. Aided by the availability of the MDD diagnosis variable in our data sets, we found that a score of 4 on the latent depression scale corresponded to a level of depression associated with a 0.85 predicted probability of having MDD in a logistic regression of MDD on depression scale scores.

Using this metric, we plotted the estimated depression distributions for all contributing data sets, which collectively span the entire range of the underlying latent depressive continuum (with nonclinical samples falling at the lower end of the scale and samples with more seriously depressed participants falling at the higher end; see Figure 1). Such breadth allowed us to assess the relation of symptom to depression across the entire range of the latent variable. More important, such heterogeneity ensures greater generalizability compared with most single-sample investigations.

Next, we conducted differential item function (DIF) tests to detect noninvariance of item parameters across the studies.[3] Of the tested items, we found no significant differences in the item characteristic curves, providing evidence of invariance across samples despite the use of different interviewers and different versions of the K–SADS. To the extent supported by the statistical results, the lack of DIF shows that our conversion of all K–SADS measures to 3-point scales yielded psychometrically equivalent metrics, thereby paving the way for tests of our more substantive hypotheses.

As shown in Figure 2, the study-specific standard error of measurement (SEM) curves convey the precision of the K–SADS at all points along the latent depression continuum. Particularly noteworthy is that the curves are horizontally aligned with one another, revealing that for all of the studies' scores from the K–SADS measure of depression were most reliable between scores of 3.1 and 5.6 on the latent depression variable. Between these values, all studies had small SEMs, ranging from 0.26 to 0.50 on the *y* axis. Also important is the fact that this "high-reliability window" contains the value of 5.0 on the *x* axis, the approximate threshold for an MDD diagnosis. At lower and higher levels of depression, the K–SADS symptom scores begin to provide a less reliable index of depression severity, as indicated by the upward curves of the SEM lines. For people with fewer than two symptoms or more than seven, the SEMs begin to exceed 1.0 on the *y* axis.[4]

Overview of main results. The relation of each symptom and the various K–SADS response options are represented by a set of response curves. As shown in the example curves in Figure 3, each symptom has three curves. The descending curve on the left represents the probability of obtaining a score of 1 (i.e., symptom is absent), as a function of the latent depression level. We would expect these probabilities to drop sharply as the level of depression increases. The rising and falling curve in the middle represents the probability of a 2 (i.e., symptom is subclinical). We would expect these probabilities to be near 0 at both the low and high ends of the depression continuum. The rising curve at the right represents the probability of a 3 (i.e., symptom is present at a clinically significant level). We would expect these probabilities to rise sharply at higher levels of the latent depression variable. The point at which the descending curve reaches .50 is called *Threshold 1*, and the point at which the rising curve meets .50 is called *Threshold 2*. These reflect symptom severity. The overall steepness of these curves reflects how sharply a symptom discriminates between different levels of depression. In the hypothetical

[2] The Hyde and Essex data set and Findling data set were each divided into two data sets, as slightly different versions of the K–SADS were used for different subsets of the participants.

[3] Because there are more than a dozen groups in the analysis, the use of IRT-based likelihood ratio (IRT-LR) DIF procedure (Thissen, Steinberg, & Wainer, 1993) was too cumbersome. Instead, we relied on the more flexible and asymptotically equivalent Wald DIF test to examine the degree to which the items exhibited cross-study differences in thresholds or discrimination parameters. For anchoring, we adopted the IRT–LR DIF convention of using all items other than the studied item as the anchor set. Due to the combination of study-specific skip patterns and missing data, some items only had a few observed responses in some studies, leading to some DIF runs with nonconverged solutions. Given this limitation, we were still able to conduct DIF tests for six of the 10 symptoms (depressed mood, irritability, anhedonia, weight and appetite disturbance, sleep disturbance, and feelings of worthlessness or guilt). There was no indication of statistically significant DIF for the symptoms tested.

[4] Because some studies in our sample have much larger or smaller variability than the reference group with an assumed variance of 1.0, the study-specific SEMs can be either larger or smaller than 1.0.

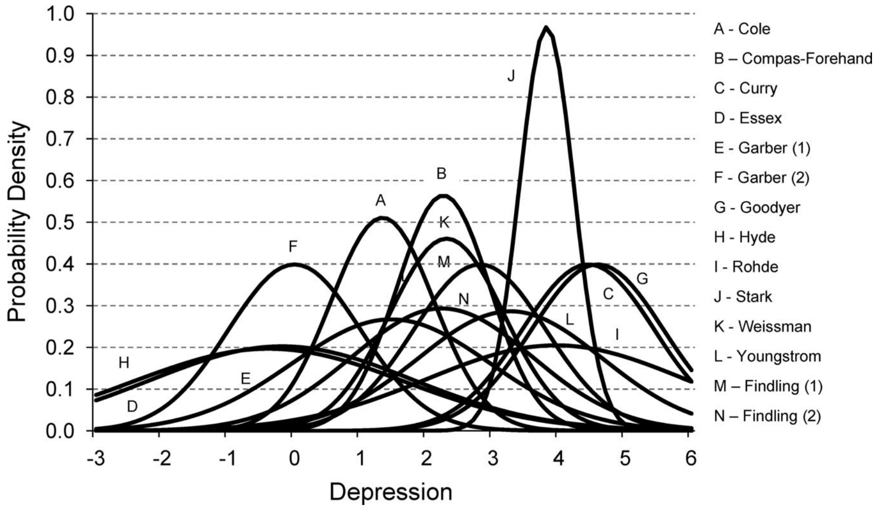

Figure 1. Distributions (probability density functions) of the contributing data sets on the latent depression variable. A = Cole (Cole et al., in press); B = Compas and Forehand (Compas et al., 2009, 2010); C = Curry (Treatment for Adolescents With Depression Study [TADS], 2003, 2005); D = Essex (Essex et al., 2006, 2009; Grabe, Hyde, & Lindberg, 2007; Mezulis, Priess, & Hyde, 2010; Priess, Lindberg, & Hyde, 2009); E = Garber 1 (Development of Depression Project [DODP], Gallerani, Garber, & Martin, 2010; Garber & Cole, 2010; Garber, Keiley, & Martin, 2002); F = Garber 2 (Parent–Child Project [PCP], Garber, Ciesla, McCauley, Diamond, & Schloredt, 2011);G = Goodyer (Goodyer et al., 2007, 2008); H = Hyde (Essex et al., 2006; 2009; Grabe, Hyde, & Lindberg, 2007; Mezulis, Priess, & Hyde, 2010; Priess, Lindberg, & Hyde, 2009); I = Rohde (N. Kaufman, Rohde, Seeley, Clarke, & Stice, 2005; Rohde, Clarke, Mace, Jorgensen, & Seeley, 2004; Rohde, Seeley, Kaufman, Clarke, & Stice, 2006); J = Stark (Fisher, 2010); K = Weissman (Pilowsky et al., 2008; Weissman, Pilowsky, & Wickramaratne, 2006); L = Youngstrom (Youngstrom et al., 2005); M & N = Findling 1 & Findling 2 (Findling et al., 2005). Note that for the present analyses, the combined data set of Essex and Hyde was split into two, and the resulting sets were labeled “Essex” and “Hyde.” The single data set for Findling also was split into two, and the resulting sets were labeled “Findling 1” and “Findling 2.”

examples of Figure 3, Panel A represents a low-severity low-discriminability symptom, Panel B represents a high-severity low-discriminability symptom, Panel C represents a low-severity high-discriminability symptom, and Panel D represents a high-severity high-discriminability symptom. Response curves for the actual symptoms appear in Figure 4, and the associated symptom threshold and discrimination parameters are the focus of the next sections.

Question 1: Are some depressive symptoms reflective of more severe depression than others? We estimated the severity thresholds for each symptom. Then we used the symptom parameter covariance matrix, produced by IRTPRO with a supplemented EM algorithm (Cai, 2008), to compute the standard errors (*SEs*) around the severity threshold estimates. With this information, we determined the rank order of the depressive symptom severities. Table 4 contains estimates of Thresholds 1 and 2 and their *SE*s. The final column of Table 4 indicates the rank order of the symptom severities, based on the second set of threshold estimates.

According to these data, clinically significant concentration problems, feelings of worthlessness or guilt, and sleep disturbance emerge at the lowest levels of depression severity, followed by problems related to depressed mood, fatigue or lack of energy, irritability, and anhedonia. At still higher levels of depression severity, psychomotor agitation or retardation, weight or appetite disturbance, and suicidal ideation or attempts emerge—with each signaling a significantly higher level of depression severity.

These results raised the possibility that concentration problems, feelings of worthlessness or guilt, and sleep disturbance might serve as a better screening cluster than depressed mood, irritability, and anhedonia (the symptoms used as screeners in some applications of the K–SADS). Consequently, we compared sensitivity and specificity analyses for the two symptom clusters. Using *DSM–IV–TR* diagnosis of MDD as the criteria, however, would bias these results in favor of the conventional screeners, as *DSM–IV–TR* requires at least one of these three symptoms for an MDD diagnosis. Instead, we used number of symptoms as the criterion. As shown in Table 5, the unconventional screeners have slightly better sensitivity than the conventional screeners. With the illness criterion set at five or more MDD symptoms, the unconventional screeners would catch 99.3%–98.6% = 0.7% more cases than would the conventional criteria. In the current data set, this translated into eight more cases. Of course, this advantage comes at the cost of lower specificity. With the illness criterion again set at five or more MDD symptoms, the conventional screeners would have correctly categorized 78.6%–70.6% = 8.0% more of people who did not have the illness, compared with the unconventional screeners. In the current data set, this translated into 114 more cases.

Question 2: Are some symptoms more discriminating indicators of depression than others? To estimate the strength of relation between each symptom and the underlying latent variable, we examined the discrimination parameters and factor loadings for each symptom (see Table 6). The item discrimination parameters can be

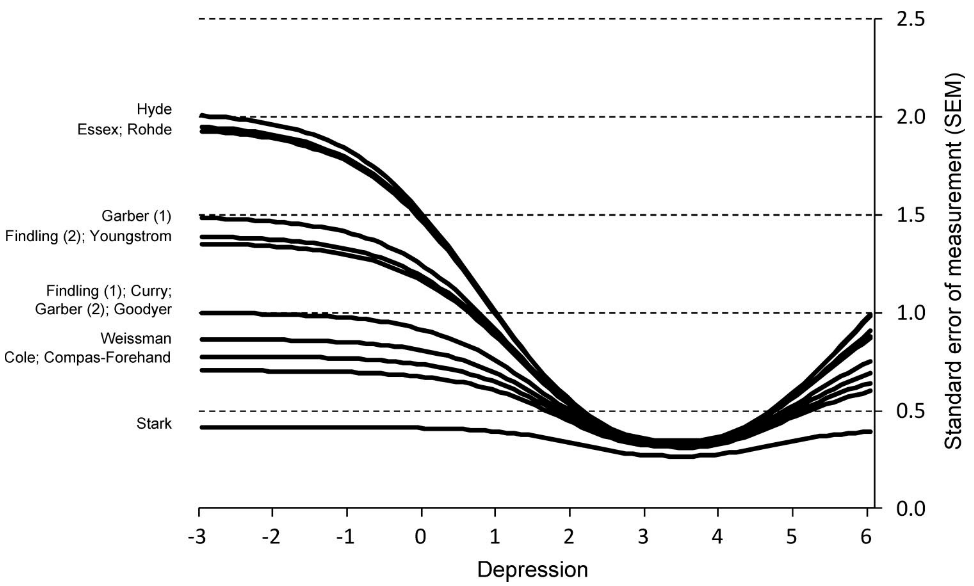

Figure 2. Standard error of measurement and Fisher information curves for all contributing studies. Hyde and Essex (Essex et al., 2006; 2009; Grabe, Hyde, & Lindberg, 2007; Mezulis, Priess, & Hyde, 2010; Priess, Lindberg, & Hyde, 2009); Rohde (N. Kaufman, Rohde, Seeley, Clarke, & Stice, 2005; Rohde, Clarke, Mace, Jorgensen, & Seeley, 2004; Rohde, Seeley, Kaufman, Clarke, & Stice, 2006); Garber 1 (Development of Depression Project [DODP], Gallerani, Garber, & Martin, 2010; Garber & Cole, 2010; Garber, Keiley, & Martin, 2002); Findling 2 (Findling et al., 2005); Youngstrom (Youngstrom et al., 2005); Findling 1 (Findling et al., 2005); Curry (Treatment for Adolescents With Depression Study [TADS], 2003, 2005); Garber 2 (Parent–Child Project [PCP], Garber, Ciesla, McCauley, Diamond, & Schloredt, 2011); Goodyer (Goodyer et al., 2007, 2008); Weissman (Pilowsky et al., 2008; Weissman, Pilowsky, & Wickramaratne, 2006); Cole (Cole et al., in press); Compas and Forehand (Compas et al., 2009, 2010); Stark (Fisher, 2010. Note that for present analyses, the combined data set of Essex and Hyde was split into two, and the resulting sets were labeled "Essex" and "Hyde." The single data set for Findling also was split into two, and the resulting sets were labeled "Findling 1" and "Findling 2."

interpreted as logistic regression slopes, or log odds-ratios. When we examined these parameters and their associated factor loadings (using conversion formulae in Wirth & Edwards, 2007), we found that all of the K–SADS items are highly discriminating indicators of depression. Even the smallest slope (suicidal ideation) is associated with depression at an odds ratio of 2.36. Examination of the overlap (and the gaps) between the confidence intervals around the slopes revealed that some symptom indicators are more discriminating than others. Depressed mood and anhedonia were the most discriminating indicators. The next most discriminating set of indicators included fatigue or lack of energy, irritability, and concentration problems. The third most discriminating set consisted of sleep disturbance, feelings of worthless and guilt, psychomotor agitation or retardation, followed by weight or appetite disturbance. The least discriminating symptom was suicidal ideation.

Question 3: How much more information can be gleaned from K–SADS interview data using IRT-based estimates of depression? One way to address this question is to compare four indices of depression severity. First was the *raw symptom count* (simply the number of *DSM*– based symptoms of depression that were coded as present). Second was the *raw symptom sum* (the raw sum of the 10 symptom variables, each on a 3-point scale). Third was called *IRT-2*, an IRT-based expected a posteriori (EAP) index based on the two-parameter logistic (2-PL) model with *only two levels* of information about presence or absence of the symptoms. And the fourth was called *IRT-3*, an IRT-based EAP index based on the graded model utilizing *all three levels* of severity for each symptom. We made this comparison by estimating the SEM for each index at varying levels of the latent depression variable. We estimated the SEM curves for the two IRT-based indices using the posterior standard deviations of the scale scores (Thissen & Wainer, 2001). We estimated crude SEM curves for the two non-IRT indices by applying the formula, $SEM = \sqrt{(1 - reliability)}$, where reliability was Cronbach's alpha for the selected index computed repeatedly for subsamples representing a sliding 2-*SD*-wide window on the latent depression variable.[5]

The four SEM curves are depicted in Figure 5. At any given level of the latent variable (i.e., various points along the *x* axis), a smaller SEM signifies greater measurement fidelity. Visual exam-

[5] The availability of the IRT scale scores, as realizations of the "true scores" of the underlying depression latent variable, enabled us to make the comparison between the reliability of raw symptom sums or counts and the reliability of the IRT scale scores. Each IRT scale score, whether IRT-2 or IRT-3, had an associated standard error of measurement. As for the raw symptom sums or counts, we calculated their reliability by treating the symptoms as observed variables in a scale and utilized a traditional summed-score-based internal consistency reliability estimator (Kuder–Richardson Formula 21 or Cronbach's alpha). The curves in Figure 5 for symptom sums or counts were smoothed to eliminate the effect of distributional discontinuities.

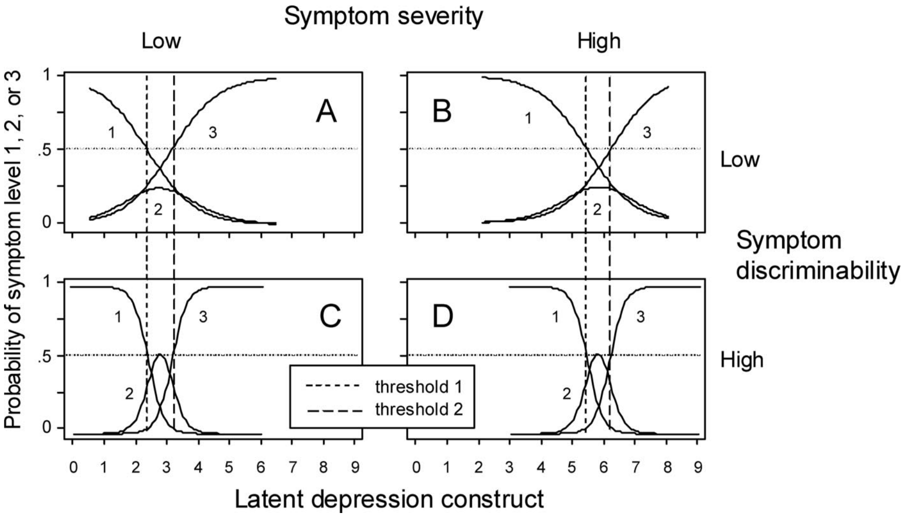

Figure 3. Hypothetical item response curves, depicting symptoms with low vs. high severity and low vs. high discriminability (1 = *symptom is absent*, 2 = *symptom is present at a subclinical level*, and 3 = *symptom is present at a clinical level*). Panel A represents a low-severity, low-discriminability symptom; Panel B represents a high-severity, low-discriminability symptom; Panel C represents a low-severity, high-discriminability symptom; and Panel D represents a high-severity, high-discriminability symptom

ination of this figure revealed two important findings. First, both of the IRT-based indices had lower SEMs than both of the non-IRT indices at virtually all levels of the latent depression variable. That is, using IRT-derived information about symptom severity and discriminability substantially enhanced precision in the measurement of depression severity. Second, both of the indices that included information about subclinical levels of depressive symptoms (i.e., the raw symptom sum and the IRT-3) were superior to both of the indices that did not include such information (i.e., raw symptom count and IRT-2). That is, both the *symptom sum* index and the *IRT-3* index had lower SEMs than the symptom count and IRT-2 index, respectively, especially at lower levels of the latent depression variable.

A second way to address this question is to examine the amount of information that is lost when one uses more conventional non-IRT-based indices of depression severity. A simple symptom count does not take into consideration the fact that some symptoms reflect greater depression severity than others. One can visualize the degree to which this is true by examining histograms depicting the range of IRT-based latent depression scores at each level of a more conventional symptom-count variable (see Figure 6). For people with a raw symptom count of 1, latent depression scores ranged from 0.7 to 3.6. For people with a raw count of 8, latent depression ranged from 4.1 to 5.8 (with an *SD* = 1.28 for the latent depression variable). This means that the variability of latent depression scores spanned approximately 1 to 2 *SD*s at each whole number value of the raw symptom count. That is, the raw symptom count gives identical scores to people with highly discrepant levels of latent depression—a process that results in a substantial loss of information.

Discussion

Four major findings about the K–SADS and depressive symptoms in children and adolescents emerged from this study. First, our K–SADS depression data were remarkably unidimensional. Second, some symptoms of depression emerged at relatively mild levels of the disorder; others emerged when depression was much more severe. Third, in children and adolescents, all K–SADS symptoms of depression were strongly associated with depression. And fourth, higher fidelity and better coverage of the construct derived from assessment algorithms in which IRT-based estimates of symptom severity and discriminability were taken into account and information about subclinical levels of symptom severity were utilized. These findings have important clinical and theoretical implications.

Our first major finding was that a very strong single latent variable emerged from our K–SADS data on symptoms of depression. Our confirmatory factor analysis showed that loadings for the 10 symptoms were strong, ranging from 0.95 (depressed mood) to 0.73 (suicide). This factor accounted for 75.4% of the covariance among the 10 symptoms. The fact that no evidence of secondary factors emerged (not even nuisance factors) is unusual for measures of depression; however, most measures of depression are questionnaires in which many symptoms are represented by multiple items. For example, the Children's Depression Inventory (Kovacs, 1985) contains three mood items, two anhedonia items, two guilt items, four self-esteem items, and so on. This creates a complex structure with a number of small factors caused by parcels of item content (Cole, Hoffman, Tram, & Maxwell, 2000). Indeed, when such measures are exceptionally unidimensional, one begins to wonder whether the items are too similar to one another, causing the underlying factor to be overly narrow. In the K–SADS, interviewers also ask multiple questions about each symptom, but then they aggregate each cluster of questions into a single appraisal about a particular symptom. This procedure greatly reduces the likelihood that nuisance factors will emerge. As the content of each item is highly distinctive (depressed mood, appetite disturbance, sleep disturbance, suicide, psychomotor ag-

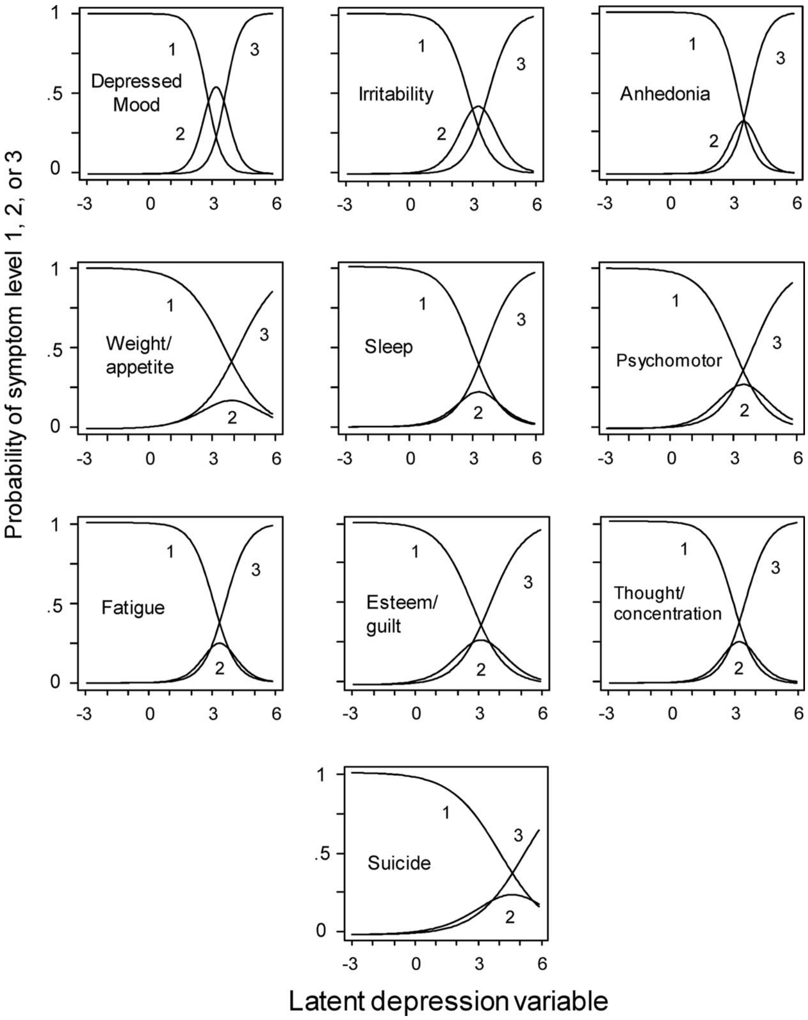

Figure 4. Item response curves for each symptom, where 1 = *symptom is absent*, 2 = *symptom is present at a subclinical level*, and 3 = *symptom is present at a clinical level.*

itation or retardation, irritability, fatigue or lack of energy, guilt or low self-esteem, concentration problems, and anhedonia), the resulting factor is anything but narrow. Given the strong, prima facie, one-to-one correspondence of K–SADS depression items to *DSM–IV–TR* depression symptoms, the emergence of a strong single factor suggests that the core symptoms of depression correlate with one another only because of a single underlying dimension of psychopathology, arguably depression.

Second, some *DSM–IV–TR* symptoms are present at significantly lower levels of depression severity than are others. At relatively low levels of the latent dimension (below the threshold for a diagnosis of MDD), clinically significant symptoms of concentration problems, feelings of worthlessness or guilt, and sleep disturbance were evident. At slightly higher levels of the latent variable (and still below the MDD threshold), symptoms of depressed mood, fatigue, irritability, and anhedonia were evident. At still higher levels of depression (and above the MDD cutoff), psychomotor agitation or retardation, weight or appetite disturbance, and suicidal ideation or attempts became increasingly likely, with each reflecting a clinically and statistically significant increase in severity on the latent variable.

Our expectation that the required symptoms of MDD (depressed mood, irritability, or anhedonia) would emerge at the lowest levels of the latent variable was not confirmed. In children and adolescents, concentration problems were evident at significantly lower levels of depression than were all three affective symptoms. Feelings of worthlessness or guilt and sleep disturbance were not significantly different from concentration problems. Taken together, these results suggest that concentration problems, feelings of worthlessness or guilt, and sleep disturbance may represent early warning signs for MDD. This possibility, however, would not seem to warrant changing the K–SADS screening criteria, as the relatively small (0.7%) gain in sensitivity comes at a much larger (8.0%) loss of specificity.

In a related vein, our results also showed that the occurrence of some symptoms signifies a much greater level of depression severity than does the occurrence of other symptoms. For example, the occurrence of feelings of worthlessness or guilt or disturbance

Table 4
Symptom Severity Parameter Estimates and Standard Errors

Symptom	Threshold 1		Threshold 2		Rank order of Threshold 2 estimates[b]
	Estimate	*SE*	Estimate[a]	*SE*	
Concentration disturbance	3.91	0.16	4.48_{a}	0.17	1
Feeling of worthlessness or guilt	3.66	0.15	4.55_{ab}	0.17	2
Sleep disturbance	3.93	0.16	4.59_{ab}	0.17	3
Depressed mood	3.76	0.15	4.65_{bc}	0.17	4
Fatigue or lack of energy	4.12	0.16	4.68_{bcd}	0.17	5
Irritability	3.80	0.15	4.79_{cd}	0.18	6
Anhedonia	4.27	0.17	4.84_{d}	0.18	7
Psychomotor disturbance	4.12	0.17	5.03_{e}	0.18	8
Weight or appetite disturbance	4.65	0.18	5.33_{f}	0.20	9
Suicide	5.08	0.19	6.24_{g}	0.23	10

Note. All threshold estimates are on the 0–9 latent depression metric. *SE* = standard error.
[a] Threshold estimates with identical subscripts are not significantly different from each other at family-wise $\alpha = .05$ with a Bonferroni correction ($z > 3.06$). [b] We focused on Threshold 2 because it represented distinction between subclinical and clinical level of symptom severity.

of sleep patterns represents a very small increase in depression severity over-and-above concentration problems, whereas the presence of psychomotor agitation or retardation, weight and appetite problems, or suicidal ideation or attempts represents substantially higher levels of severity. To our knowledge, no K–SADS measurement algorithm makes use of this kind of information, which could substantially enhance the fidelity of depression severity assessments.

Third, all *DSM–IV–TR* symptoms were strong indicators of depression in children and adolescents; however, some symptoms were more strongly related to the depression factor than others. Depressed mood was by far the strongest indicator, such that a 1-point change in the latent variable was associated with a 15-fold increase in the probability of the symptom. Anhedonia was the next most discriminating symptom, followed by fatigue or lack of energy, irritability, and concentration problems. Suicidal ideation or attempt was the least discriminating symptom. This kind of information can be used to enhance the measurement of depression severity (Weiss, 1982, 1985).

We found that an interesting trade-off appears to exist between severity and discriminability of depressive symptoms as indicators of depression, with the less discriminating items emerging at higher levels of depression severity. The correlation between severity and discriminability estimates was −0.64. For example, suicidal ideation or attempts, weight or appetite disturbance, and psychomotor agitation or retardation were among the most severe yet least discriminating symptoms. Conversely, depressed mood and concentration problems were among the less severe but more discriminating symptoms. In an ideal method of measurement, both severity and discriminability would be taken into consideration.

Fourth, to do this, we constructed two IRT-based K–SADS indices of depression severity; one was based on just the presence or absence of symptoms, whereas the other utilized information about subclinical symptoms as well. Both indices outperformed more conventional scoring methods that were simply based on symptom counts or summed scores. Psychometric comparisons revealed that scores from the IRT-based measures were more reliable and had lower *SEs*, especially in the moderate to severe range of depression. Furthermore, utilizing subclinical symptom information extended these psychometric advantages further into the mild range of depression. In other words, using IRT methods

Table 5
Sensitivity and Specificity Analyses for Two Different Screening Tests: Conventional Versus Unconventional

Screening test	Criterion: How many MDD symptoms present							
	≥1	≥2	≥3	≥4	≥5[a]	≥6	≥7	≥8
Sensitivity								
Conventional			.942	.973	**.986**	.995	.996	1.000
Unconventional			.980	.991	**.993**	1.000	1.000	1.000
Specificity								
Conventional	1.000	.942	.894	.843	**.786**			
Unconventional	1.000	.915	.837	.768	**.706**			

Note. Conventional = depressed mood, irritability, or anhedonia; unconventional = concentration problems, feelings of worthlessness and guilt, or sleep disturbance; MDD = major depressive disorder; sensitivity = proportion of people reaching criterion symptom count who had a positive outcome on the screening test; specificity = proportion of people not reaching criterion symptom count who had a negative outcome on the screening test.
[a] *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.) criteria for MDD requires at least five symptoms. Bold type indicates values for these criteria.

Table 6
Estimates of Symptom Discrimination Parameters and Factor Loadings

Symptom	Discrimination[a]	*SE*	Factor loading	*SE*	Odds ratio
Depressed mood	2.71_a	0.15	0.94	0.05	15.03
Anhedonia	2.33_a	0.13	0.89	0.05	10.28
Fatigue	1.83_b	0.10	0.79	0.05	6.23
Irritability	1.80_b	0.09	0.78	0.05	6.05
Concentration problems	1.77_b	0.10	0.78	0.05	5.87
Sleep	1.39_c	0.08	0.67	0.05	4.01
Feelings of worthlessness or guilt	1.24_{cd}	0.07	0.62	0.05	3.46
Psychomotor disturbance	1.23_{cd}	0.07	0.61	0.05	3.42
Weight or appetite disturbance	1.03_{de}	0.06	0.54	0.05	2.80
Suicide	0.86_e	0.05	0.46	0.04	2.36

Note. *SE* = standard error.
[a] Estimates with identical subscripts are not significantly different from each other at family-wise $\alpha = .05$ with a Bonferroni correction ($z > 3.06$).

and incorporating information about subclinical symptom levels increased both the fidelity and bandwidth of measurement.

These results have two noteworthy implications. First, IRT-based increments in measurement fidelity (i.e., reduced measurement error) can readily translate into larger between- and within-group effect sizes and therefore into greater statistical power to detect treatment effects, as shown in at least one randomized treatment-control study on the effectiveness of antidepressants (Santor, Debrota, Engelhardt, & Gelwicks, 2008). Second, the IRT-based inclusion of subclinical symptom information and the resultant increased bandwidth can be especially helpful in treatment-comparison research. When one treatment is compared with another, a large part of the effect can depend upon differences that reside in the subclinical range of the dependent variable. The inclusion of even one extra response option to indicate the subclinical presence of each symptom can substantially enhance the researcher's capacity to detect a treatment difference. Whether the inclusion of even more response options could generate more power is an interesting question worthy of further investigation.

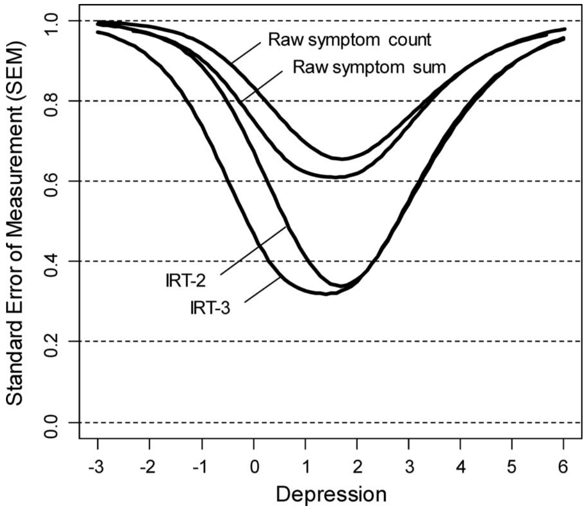

Figure 5. Standard error of measurement curves for four indices of depression severity, as a function of the latent depression variable. IRT = item response theory; IRT-2 = IRT-based expected a posteriori (EAP) index based on the 2-PL model utilizing only two levels of information about presence or absence of the symptoms.IRT-3 = IRT-based EAP index based on the graded model utilizing all three levels of severity for each symptom.

At least four shortcomings of the current study suggest avenues for future research. First, all of the IRT analyses in this study focused on data obtained by using the K–SADS. Although this measure utilizes information from multiple informants, filtered through the expertise of well-trained clinical interviewers, the K–SADS still represents only a single method for measuring depression. As such, it is possible that the strong latent variable that emerged from our analyses represents not just depression but also this method. Although semistructured clinical interviews like the K–SADS have been touted as the closest thing to a gold standard that mental health researchers have in the assessment of psychopathology (Hersen & Gross, 2008), they are not immune to method effects. Although it is unlikely that demand characteristics or interviewer bias would act similarly across all the investigative teams that contributed data to this study, it is not impossible. For example, eager to fill the quota of depressed participants in a research study, interviewers could have been positively biased in their perception of depressive symptoms. Replication of the current results with multiple, methodologically dissimilar measures of depression would mitigate these concerns.

Second, our analyses carefully established the invariance of the IRT results across the samples that contributed to the aggregate data set. This is a critical first step. It is possible, however, that the results may not be invariant across other ways of subdividing the data. Efforts are currently underway to examine ways that the relation of symptoms to the underlying depression factor may vary as a function of age, gender, and ethnicity.

Third, though we were able to use IRT methods to accomplish the cross-study linkage of latent variable scales, our results are best treated as a first step, in the absence of further evaluations of the quality of linking. Furthermore, we note that the means of the studies are spread out widely across the latent depression scale, which can lead to a deterioration of the quality of linking in the extremes.

Finally, the current study provided very strong evidence that a single underlying factor underlies the 10 symptoms of depression as assessed by the K–SADS. It is possible, however, that this

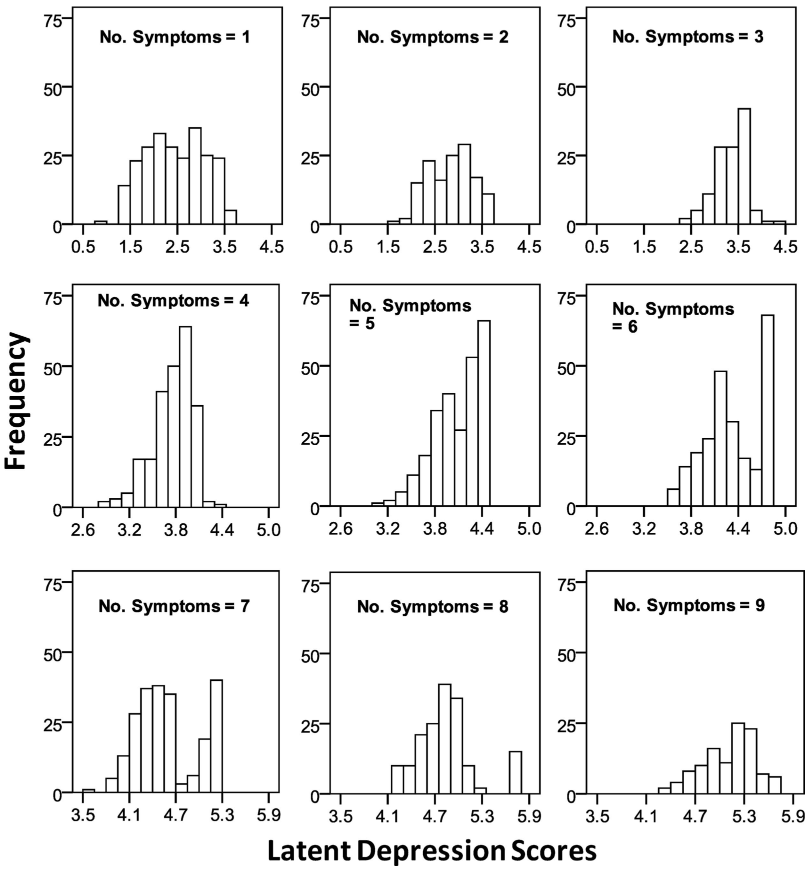

Figure 6. Histograms of latent depression levels at each level of symptom count index of depression based on the Kiddie Schedule of Affective Disorders and Schizophrenia for School-Aged Children.

unidimensionality depends upon the level at which the symptoms of depression are examined. We focused on symptom clusters, as recommended in the *DSM–IV–TR* for the diagnosis of MDD. Specific examples include negative self-perceptions (which consist of low self-esteem and guilt feeling), irritability and anger, sleep disturbance (hypersomnia and insomnia), psychomotor symptoms (agitation and retardation), and weight or appetite disturbance (increase and decrease). Examination of the disaggregated symptoms could reveal evidence of one or more other dimensions.

References

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). *DSM* criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine, 35,* 475–487. doi:10.1017/S0033291704003563

Ambrosini, P. (2000). Historical development and present status of the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K–SADS): Research psychiatric diagnostic interviews for children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 39,* 49–58. doi:10.1097/00004583-200001000-00016

Ambrosini, P., & Dixon, J. (1996). *Schedule for Affective Disorders and Schizophrenia for School-Aged Children—Present Version, Version IV–Revised (K–SADS–IV–R).* Unpublished instrument, Eastern Pennsylvania Psychiatric Institute, Medical College of Pennsylvania, Philadelphia, PA.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Bedi, R. P., Maraun, M. D., & Chrisjohn, R. D. (2001). A multisample item response theory analysis of the Beck Depression Inventory. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement, 33,* 176–187. doi:10.1037/h0087139

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459. doi:10.1007/BF02293801

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280. doi:10.1177/014662168801200305

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61,* 309–329. doi:10.1348/000711007X249603

Cai, L., du Toit, S. H. C., & Thissen, D. (in press). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling.* Chicago, IL: Scientific Software International.

Cassano, G. B., Benvenuti, A., Miniati, M., Calugi, S., Mula, M., Maggi, L., . . . Frank, E. (2009). The factor structure of lifetime depressive spectrum in patients with unipolar depression. *Journal of Affective Disorders, 115,* 87–99. doi:10.1016/j.jad.2008.09.006

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265–289.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100,* 316–336. doi:10.1037/0021-843X.100.3.316

Cole, D. A., Hoffman, K., Tram, J. M., & Maxwell, S. E. (2000). Structural differences in parent and child reports of children's symptoms of depression and anxiety. *Psychological Assessment, 12,* 174–185. doi: 10.1037/1040-3590.12.2.174

Cole, D. A., Jacquez, F. M., LaGrange, B., Pineda, A. Q., Truss, A. E., Weitlauf, A. S., . . . Dufton, L. (2010). A longitudinal study of cognitive risks for depressive symptoms in children and young adolescents. *Journal of Early Adolescence*. Advance online publication. doi: 10.1177/0272431610376248

Compas, B. E., Champion, J. E., Forehand, R., Cole, D. A., Reeslund, K. L., Fear, J., . . . Roberts, L. (2010). Coping and parenting: Mediators of 12-month outcomes of a family group cognitive-behavioral preventive intervention with families of depressed parents. *Journal of Consulting and Clinical Psychology, 78,* 623–634. doi:10.1037/a0020459

Compas, B. E., Forehand, R., Keller, G., Champion, J. E., Rakow, A., Reeslund, K. L., . . . Cole, D. A. (2009). Randomized controlled trial of a family cognitive–behavioral preventive intervention for children of depressed parents. *Journal of Consulting and Clinical Psychology, 77,* 1007–1020. doi:10.1037/a0016930

Essex, M. J., Kraemer, H. C., Armstrong, J. M., Boyce, W. T., Goldsmith, H. H., Klein, M. H., . . . Kupfer, D. J. (2006). Exploring risk factors for the emergence of children's mental health problems. *Archives of General Psychiatry, 63,* 1246–1256. doi:10.1001/archpsyc.63.11.1246

Essex, M. J., Kraemer, H. C., Slattery, M. J., Burk, L. R., Boyce, W. T., Woodward, H. R., . . . & Kupfer, D. J. (2009). Screening for childhood mental health problems: Outcomes and early identification. *Journal of Child Psychology and Psychiatry, 50,* 562–570. doi:10.1111/j.1469-7610.2008.02015.x

Findling, R. L., Youngstrom, E. A., McNamara, N. K., Stansbrey, R. J., Demeter, C. A., Bedoya, D., . . . Calabrese, J. R. (2005). Early symptoms of mania and the role of parental risk. *Bipolar Disorders, 7,* 623–634. doi:10.1111/j.1399-5618.2005.00260.x

Fisher, M. E. (2010). *Examining sudden gains during cognitive-behavioral therapy for depressed 9 to 13 year old girls* [Unpublished doctoral dissertation]. The University of Texas at Austin, TX.

Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. B., & Klapp, B. F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research, 18,* 23–36. doi:10.1002/mpr.274

Gallerani, C. M., Garber, J., & Martin, N. C. (2010). The temporal relation between depression and comorbid psychopathy in adolescents at varied risk for depression. *Journal of Child Psychology and Psychiatry, 51,* 242–249.

Garber, J., & Cole, D. A. (2010). Intergenerational transmission of depression: A launch and grow model of change across adolescence. *Development and Psychopathology, 22,* 819–830.

Garber, J., Ciesla, J. A., McCauley, E., Diamond, G. S., & Schloredt, K. A. (2011). Remission of depression in parents: Links to healthy functioning in their children. *Child Development, 82,* 244–261.

Garber, J., Keiley, M. K., & Martin, N. C. (2002). Developmental trajectories of adolescents' depressive symptoms: Predictors of change. *Journal of Consulting and Clinical Psychology, 70,* 79–95. doi:10.1037/0022-006X.70.1.79

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E.(2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry, 4,* 13. doi:10.1186/1471-244X-4-13

Geller, B., Zimerman, B., Williams, M., Bolhofner, K., Craney, J. L., Delbello, M. P., & Soutullo, C. (2001). Reliability of the Washington University in St. Louis Kiddie Schedule for Affective Disorders and Schizophrenia (WASH-U-KSADS) mania and rapid cycling sections. *Journal of the American Academy of Child & Adolescent Psychiatry, 40,* 450–455. doi:10.1097/00004583-200104000-00014

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59,* 361–368. doi:10.1176/appi.ps.59.4.361

Goodyer, I., Dubicka, B., Wilkinson, P., Kelvin, R., Roberts, C., Byford, S., . . . Harrington, R. (2007). Selective serotonin reuptake inhibitors (SSRIs) and routine specialist care with and without cognitive behavior therapy in adolescents with major depression: Randomised controlled trial. *British Medical Journal* 335, 142. doi:10.1136/bmj.39224.494340.55

Goodyer, I., Dubicka, B., Wilkinson, P., Kelvin, R., Roberts, C., Byford, S., . . . Harrington, R. (2008). A randomized controlled trial of cognitive behavior therapy in adolescents with major depression treated by selective serotonin reuptake inhibitors. The ADAPT trial. *Health Technology Assessment, 12,* ix–60.

Grabe, S., Hyde, J., & Lindberg, S. (2007). Body objectification and depression in adolescents: The role of gender, shame, and rumination. *Psychology of Women Quarterly, 31,* 164–175. doi:10.1111/j.1471-6402.2007.00350.x

Hersen, M., & Gross, A. M. (Eds.). (2008). *Handbook of clinical psychology*: *Vol.* 2. *Children and adolescents*. Hoboken, NJ: Wiley.

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., . . . Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School Age Children–Present and Lifetime Version (K–SADS–PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry, 36,* 980–988. doi:10.1097/00004583-199707000-00021

Kaufman, J., Birmaher, B., Brent, D., Rao, U., & Ryan, N. (1996). *Kiddie SADS– Present and Lifetime Version (K–SADS–PL).* Unpublished instrument, Western Psychiatric Institute and Clinics, University of Pittsburgh School of Medicine, PA.

Kaufman, N. K., Rohde, P., Seeley, J. R., Clarke, G. N., & Stice, E. (2005). Potential mediators of cognitive–behavioral therapy for adolescents with comorbid major depressive and conduct disorder. *Journal of Consulting and Clinical Psychology, 73,* 38–46. doi:10.1037/0022-006X.73.1.38

Kovacs, M. (1985). The Children's Depression Inventory (CDI). *Psychopharmacology Bulletin, 21,* 995–998.

Lonigan, C. J., Carey, M. P., & Finch, A. J. (1994). Anxiety and depression in children and adolescents: Negative affectivity and the utility of self-reports. *Journal of Consulting and Clinical Psychology, 62,* 1000–1008. doi:10.1037/0022-006X.62.5.1000

Mezulis, A. H., Priess, H. A., & Hyde, J. S. (2010). Rumination mediates the relationship between infant temperament and adolescent depressive symptoms. *Depression Research and Treatment*. Advance online publication. doi:10.1155/2011/487873

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64. doi:10.1177/01466216000241003

Orvaschel, H. (1994). *Schedule for Affective Disorders and Schizophrenia for School-Age Children—Epidemiological Version* (5th ed.). Ft. Lauderdale, FL: Nova Southeastern University.

Pilowsky, D. J., Wickramaratne, P., Talati, A., Tang, M., Hughes, C. W., Garber, J., . . . Weissman, M. M. (2008). Children of depressed mothers 1 year after the initiation of maternal treatment: Findings from the STAR*D-Child Study. *American Journal of Psychiatry, 165,* 1136–1147. doi:10.1176/appi.ajp.2008.07081286

Priess, H., Lindberg, S., & Hyde, J. S. (2009). Adolescent gender-role identity and mental health: Gender intensification revisited. *Child Development, 80,* 1531-1544.

Reeve, B. B., Burke, L. B., Chiang, Y.-P.., Clauser, S. B., Colpe, L. J., Elias, J. W., . . .Werner, E. M. (2007). Enhancing measurement in health outcomes research supported by agencies within the U.S. Department of Health and Human Services. *Quality of Life Research, 16*(Suppl.1), 175–186. doi:10.1007/s11136-007-9190-8

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . David, C., on behalf of the PROMIS Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(5, Suppl. 1), S22–S31. doi:10.1097/01.mlr.0000250483.85507.04

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5,* 27–48. doi:10.1146/annurev.clinpsy.032408.153553

Rohde, P., Clarke, G., Mace, D., Jorgensen, J., & Seeley, J. (2004). An efficacy/effectiveness study of cognitive-behavioral treatment for adolescents with comorbid major depression and conduct disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 43,* 660–668. doi:10.1097/01.chi.0000121067.29744.41

Rohde, P., Seeley, J., Kaufman, N., Clarke, G., & Stice, E. (2006). Predicting time to recovery among depressed adolescents treated in two psychosocial group interventions. *Journal of Consulting and Clinical Psychology, 74,* 80–88. doi:10.1037/0022-006X.74.1.80

Ryan, N. D., Puig-Antich, J., Ambrosini, P., Rabinovich, H., Robinson, D., Nelson, B., . . . Twomey, J. (1987). The clinical picture of major depression in children and adolescents. *Archives of General Psychiatry, 44,* 854–861.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4), 100–114.

Santor, D. A., Debrota, D., Engelhardt, N., & Gelwicks, S. (2008). Optimizing the ability of the Hamilton Depression Rating Scale to discriminate across levels of severity and between antidepressants and placebos. *Depression and Anxiety, 25,* 774–786. doi:10.1002/da.20351

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

Simon, G. E., & Von Korff, M. (2006). Medical co-morbidity and validity of *DSM-IV* depression criteria. *Psychological Medicine, 36,* 27–36. doi:10.1017/S0033291705006136

Small, D. M., Simons, A. D., Yovanoff, P., Silva, S. G., Lewis, C. C., Murakami, J. L., & March, J. (2008). Depressed adolescents and comorbid psychiatric disorders: Are there differences in the presentation of depression? *Journal of Abnormal Child Psychology, 36,* 1015–1028. doi:10.1007/s10802-008-9237-5

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Mahwah, NJ: Erlbaum.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

Treatment for Adolescents With Depression Study (TADS) Team. (2003). The Treatment for Adolescents With Depression Study (TADS): Rationale, design, and methods. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 531–542. doi:10.1097/01.CHI.0000046839.90931.0D

Treatment for Adolescents With Depression Study (TADS) Team. (2005). The Treatment for Adolescents With Depression Study (TADS): Demographic and clinical characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry, 44,* 28–40. doi:10.1097/01.chi.0000145807.09027.82

Waller, N. G., Compas, B. E., Hollon, S. D., & Beckjord, E. (2005). Measurement of depressive symptoms in women with breast cancer and women with clinical depression: A differential item functioning analysis. *Journal of Clinical Psychology in Medical Settings, 12,* 127–141. doi:10.1007/s10880-005-3273-x

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive theory. *Applied Psychological Measurement, 6,* 473–492. doi:10.1177/014662168200600408

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774–789. doi:10.1037/0022-006X.53.6.774

Weissman, M. M., Pilowsky, D. J., Wickramaratne, P. J., Talati, A., Wisniewski, S. R., Fava, M., . . . Rush, A. J., for the STAR*D Child Team. (2006). Remissions in maternal depression and child psychopathology: A STAR*D-Child report. *Journal of the American Medical Association, 295,* 1389–1398. doi:10.1001/jama.295.12.1389

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79. doi:10.1037/1082-989X.12.1.58

Youngstrom, E., Meyers, O., Demeter, C., Youngstrom, J., Morello, L., Piiparinen, R., . . . Findling, R. L. (2005). Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders, 7,* 507–517. doi:10.1111/j.1399-5618.2005.00269.x

Received March 9, 2010
Revision received February 7, 2011
Accepted February 7, 2011 ■