Psychological Assessment
2012, Vol. 24, No. 2, 341–351

1040-3590/11/$12.00 DOI: 10.1037/a0025617

Portability of a Screener for Pediatric Bipolar Disorder to a Diverse Setting

Andrew J. Freeman and Eric A. Youngstrom
University of North Carolina at Chapel Hill

Thomas W. Frazier
Cleveland Clinic Foundation

Jennifer Kogos Youngstrom
University of North Carolina at Chapel Hill

Christine Demeter and Robert L. Findling
Case Western Reserve University

Robust screening measures that perform well in different populations could help improve the accuracy of diagnosis of pediatric bipolar disorder. Changes in sampling could influence the performance of items and potentially influence total scores enough to alter the predictive utility of scores. Additionally, creating a brief version of a measure by extracting items from a longer scale might cause differential functioning due to context effects. The authors of current study examined both sampling and context effects of a brief measure of pediatric mania. Caregivers of 813 youths completed the parent-reported version of the General Behavior Inventory (PGBI) at an academic medical center sample enriched for mood disorders. Caregivers of 481 youths completed the PGBI at a community mental health center. Caregivers of 799 youths completed 10 items extracted from the PGBI at a community setting. Caregivers of 159 youths completed both versions of the PGBI and a semistructured diagnostic interview. Differential item functioning indicated that across samples some items functioned differently; however, total observed scores were similar across all levels of mania. Receiver operating characteristic analysis indicated that the 10 extracted items discriminated bipolar disorder from nonbipolar behavior as well as when the items were embedded within the full measure. Findings suggest that the extracted items perform similarly to the embedded items in the community setting. Measurement properties appear sufficiently robust across settings to support clinical applications.

Keywords: bipolar disorder, sensitivity, specificity, differential item functioning

Clinic visits associated with pediatric bipolar disorder (PBD) have increased fortyfold in the last decade (Moreno et al., 2007). General population prevalence estimates suggest that up to 1.8% of youths are affected with bipolar spectrum disorders, compared with traditionally held views that bipolar disorder is an adult diagnosis and extremely rare in childhood and adolescence (Van Meter, Moreira, & Youngstrom, 2011). However, clinical and research diagnoses of mood disorders in both youths and adults show substantial disagreement, suggesting that clinicians and researchers might be focused on different symptom presentations (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). As a result, substantial controversy surrounds the diagnosis of PBD. There is a clear need for evidence-based assessment approaches to PBD.

Accurate assessment of PBD relies on assessing the frequency, intensity, number, and duration of hypomanic and manic symptoms (Quinn & Fristad, 2004). The symptoms of hypomania and mania are identical, and the two states are differentiated by duration and intensity: Mania requires either a week of mood disturbance or psychiatric hospitalization, whereas a hypomanic episode involves more mild or moderate symptoms lasting at least 4 days (American Psychiatric Association, 2004). Both hypomanic and manic episodes in PBD are characterized by periods of time during which youths experience elevated mood, increased energy, irritability, and often also grandiosity or decreased need for sleep (Youngstrom, Birmaher, & Findling, 2008). The combination of the most severe lifetime hypomanic, manic, and depressive episodes determines the presence and subtype of bipolar disorder (American Psychiatric Association, 2004). Relative to adults, youths may experience longer episodes (Birmaher et al., 2006), and symptoms overlap with other common childhood disorders such as attention-deficit/hyperactivity disorder (ADHD; Bowring & Kovacs, 1992). Adding to the difficulty in determining the

This article was published Online First September 26, 2011.

Andrew J. Freeman, Eric A. Youngstrom, and Jennifer Kogos Youngstrom, Department of Psychology, University of North Carolina at Chapel Hill; Thomas W. Frazier, Center for Autism, Cleveland Clinic Foundation; Christine Demeter and Robert L. Findling, Department of Psychiatry, Case Western Reserve University.

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

The work was supported in part by National Institute of Mental Health Grant NIHR01 MH066647 (Principal Investigator, Eric A. Youngstrom). We thank the families who participated in this research. We thank Andrea Hussong and David Thissen for their comments during their service on the first author's master's committee.

Correspondence concerning this article should be addressed to Andrew J. Freeman, Department of Psychology, University of North Carolina at Chapel Hill, CB#3270, Chapel Hill, NC, 27599. E-mail: Andrew.Freeman@unc.edu

origins of a symptom is the fact that most self-referred treatment seeking occurs during depressive episodes (Youngstrom, Freeman, & Jenkins, 2009). Therefore, a brief accurate screening measure for (hypo)manic symptoms could increase the accuracy of PBD diagnoses (Henry, Pavuluri, Youngstrom, & Birmaher, 2008; Jenkins, Youngstrom, Washburn, & Youngstrom, 2011; Youngstrom, Frazier, Demeter, Calabrese, & Findling, 2008).

Numerous measures have been proposed in the research literature to improve the assessment of PBD because early and accurate identification may lead to more effective treatment. Measures of PBD assess the presence of hypomanic and manic symptoms because the diagnosis of bipolar disorder is differentiated from other disorders by the presence of hypomanic and manic episodes (American Psychiatric Association, 2004, 2011; see Miller, Johnson, & Eisner, 2009; Youngstrom, Mash, & Barkley, 2007, for review). Validation studies of manic symptom measures typically have compared performance by a bipolar group to the performance of healthy controls and a single comparison group such as individuals with major depression (Hirschfeld et al., 2000) or ADHD (Pavuluri, Henry, Devineni, Carbray, & Birmaher, 2006; Tillman & Geller, 2005). Changes in comorbidity patterns with overlapping symptoms—such as increases in comorbid disruptive behavior disorders—could result in measures performing more poorly (Kowatch, Youngstrom, Danielyan, & Findling, 2005; Neighbors, Jackson, Campbell, & Williams, 1989; Youngstrom & Green, 2003). For example, items such as "cries often and easily" and "mood changes quickly and drastically" displayed adequate sensitivity and specificity to bipolar I disorder in a distilled sample that excluded cases with conduct disorder or comorbid ADHD and depression, but failed to discriminate PBD from other diagnoses in a more diagnostically diverse sample (Tillman & Geller, 2005; cf. Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006). In addition, the average severity of mania may often be lower in community mental health settings than in specialty clinics. For example, the Mood Disorder Questionnaire (MDQ; Hirschfeld et al., 2000) demonstrates substantial sensitivity to bipolar I disorder; however, the MDQ displays poor sensitivity to bipolar II and bipolar spectrum disorders (Hirschfeld et al., 2003; Miller, Klugman, Berv, Rosenquist, & Ghaemi, 2004; Wagner et al., 2006; Youngstrom, Meyers, et al., 2005). These other diagnoses on the bipolar spectrum appear to be more common than bipolar I in both clinical (Birmaher et al., 2006) and community (Merikangas & Pato, 2009; Van Meter, et al., 2011) samples. Thus, existing evidence strongly suggests that measures developed in highly selected samples might not generalize to community mental health populations due to changes in clinical characteristics.

To be used in widespread screening of a diagnosis, the measure should be robust across diverse samples (Kraemer, 1992; Straus, Richardson, Glasziou, & Haynes, 2005). In one direct comparison, fewer than half of measures displayed good discrimination of PBD from other diagnoses in more clinically representative samples, with only small decreases in accuracy observed (Youngstrom et al., 2006). The parent-reported version of the General Behavior Inventory (PGBI) displayed excellent functioning in both an academic medical center and community mental health clinic. The PGBI (Youngstrom, Findling, Danielson, & Calabrese, 2001) represents an adaptation of the General Behavior Inventory (Depue, 1981) from college student self-report to caregiver reporting of youths. The target of the item query changed from self to offspring because the criteria for bipolar disorder are the same between youths and adults or among informants (American Psychiatric Association, 2004, 2011; Youngstrom, Birmaher, et al., 2008).

The PGBI displays both positive and negative attributes for the assessment of PBD. The PGBI assesses mixed symptoms, mood lability, and episodes while maintaining adequate sensitivity and specificity to bipolar spectrum diagnoses (Youngstrom et al., 2001), whereas many other measures query only about the presence of manic symptoms without mixed presentations (e.g., Pavuluri et al., 2006; Wagner et al., 2006). Mixed symptom presentation is common in youths (Kraepelin, 1921/1976; Youngstrom, Birmaher, et al., 2008). The PGBI also displays sensitivity to treatment effects.

Undesirable characteristics for widespread use of the PGBI are length (73 items) and reading level (12th grade). To decrease the burden, Youngstrom, Frazier, et al. (2008) developed the 10-item mania PGBI (PGBI–10M) by extracting the 10 items that were most discriminating between PBD and all other diagnoses at an academic medical center. The content of those 10 items stayed the same between the PGBI and PGBI–10M.

Extracting items could result in a change of response context. Context effects are traditionally defined as the interaction between the content of prior items with the current item (Schuman, Presser, & Ludwig, 1981). The content of the 73 items of the PGBI provides a general context that directly and consistently queries about mood symptoms. Thus, it is possible that item and test functioning could change as a result of the change in context. One major difference in context is that the 73 items include a separate Depression scale as well as a Hypomanic/Biphasic (i.e., "mixed") scale, whereas the 10 items comprising the PGBI–10M are drawn solely from the Hypomanic/Biphasic scale.

Item functioning is most often examined using item response theory, which is a method for examining both an item and the test's functioning on an underlying latent trait. Two parameter logistic models provide estimates of discrimination and threshold. In the context of psychopathology, the *discrimination* parameter represents the likelihood that an individual will endorse the symptom at his or her severity of mania and the *threshold* parameter represents the severity at which there is a 50% probability of endorsing this response or higher. Differential item functioning (DIF) occurs when two groups with the same estimated severity do not have the same probability of choosing identical responses (Lord, 1980). Thus, item response theory provides a framework for examining the effects of changing sampling and context on caregiver response to the PGBI and PGBI–10M.

The present study examined the extent to which psychometric properties changed when the PGBI–10M was transported into new settings. Specific aims included the following:

1. Examine differential item functioning and differential test functioning of the 10 items on the PGBI between two socioeconomic, racially and clinically distinct samples.

2. Examine the differential item functioning and differential test functioning of the extracted 10 items in the form of the PGBI–10M compared with the embedded 10 items in the form of the full PGBI in a low socioeconomic, racially and clinically diverse sample.

3. Examine the convergent and discriminant validity of the PGBI–10M when administered separately compared with the 10 items embedded within the PGBI in a low socioeconomic, racially and clinically diverse sample.

4. Examine the diagnostic efficiency of the PGBI–10M when administered in a low socioeconomic, racially and clinically diverse sample.

Method

Participants

Participants were 2,252 youths presenting at either an urban academic medical center ($n = 813$) or an urban community mental health center ($n = 1,439$) in the Midwest. The community mental health sample was an unselected case series that would be representative of youths seeking services in urban, low-income settings. The academic medical center has specialty clinics in adult and pediatric bipolar disorder and recruits cases to fill research studies. Families contacting the academic medical center before 2003 went through a phone screening and were referred to other providers if they did not meet criteria for inclusion in one of the research projects. Additionally, advertising for studies and referrals of offspring with a parent with bipolar disorder enriched the rate of bipolar disorder in the academic sample. Inclusion criteria for the current study at both sites were (a) youths were between the ages of 5 and 18 years and were seeking outpatient mental health services, (b) both the caregiver and youth provided written consent and assent, (c) both the caregiver and youth presented for the assessment, and (d) both the caregiver and youth were conversant in English. Table 1 displays the demographic characteristics of the sample divided into subgroups for analysis. Overall, participants in the community mental health sample were more likely to be African American and have no mood disorder, whereas participants at the academic medical center were more likely to be White or have bipolar I. Rates of bipolar I disorder in the community mental health clinic were (a) substantially higher than found in nonclinical community samples (Merikangas et al., 2010; Van Meter, et al., 2011), (b) similar to other published rates for similar samples (Geller et al., 2002; Youngstrom, Youngstrom, & Starr, 2005), and (c) lower than rates found in settings that treat youths with greater acuity of problems (Blader & Carlson, 2007; Pliszka, Sherman, Barrow, & Irick, 2000). The fourfold increase in the rates of bipolar spectrum diagnoses compared with bipolar I rates is consistent with epidemiological findings that indicate a fourfold increase in bipolar spectrum disorders compared with bipolar I (Lewinsohn, Klein, & Seeley, 1995; Merikangas et al., 2011).

The total sample was split into four groups: Embedded Academic (EA), Embedded Community (EC), Extracted, and Both. The EA group consisted of 813 youths and their caregivers from an academic medical center. The EC group consisted of 481 youths from the community mental health center. The primary caregivers of the EA and EC youths completed the full PGBI. The Extracted group consisted of 799 youths from the community mental center, whose parents completed the PGBI–10M only as stand-alone measure during general intake to the clinic. The both group consisted of 159 youths from the community mental health center whose parents completed both the PGBI–10M at general intake and then later completed full PGBI during an expanded research protocol (median: 8 days after intake). The Extracted group did not participate in the larger protocol, so demographic and clinical characteristics were not gathered at an individual level. Like the EC group, the Extracted group was a case series at the same clinical infrastructure so demographic and clinical features would be similar.

The academic medical center site had multiple pharmacotherapy trials open for bipolar spectrum disorders, unipolar depression,

Table 1
Demographic Characteristics of the Embedded Academic, Embedded Community, and Both Outpatient Groups

Variable	Group		
	Embedded academic ($n = 813$)	Embedded community ($n = 481$)	Both ($n = 159$)
Gender (%)			
Male	61	58	65
Female	39	42	35
Ethnicity (%)			
White	79	9	4
African American	13	83	91
Age in years (*SD*)	11.5 (3.3)	10.8 (3.4)	10.0 (3.4)
Number of diagnoses (*SD*)	2.1 (1.3)	2.7 (1.4)	2.6 (1.2)
Primary diagnosis (%)			
Bipolar I	23	3%	1%
Other bipolar spectrum	20	11%	10%
Unipolar depression	23	31%	21%
Disruptive behavior disorder without mood	23	45%	57%
All other diagnoses	11	9%	9%

Note. Demographics were not available for the extracted group. Composition should be similar to the embedded community sample, as both samples were consecutive case series at the same infrastructure. Bipolar spectrum includes bipolar II, cyclothymia, and bipolar-not otherwise specified in accordance with *DSM–IV–TR*. Primary diagnoses were hierarchically determined such that if a youth had bipolar I and comorbidity, the primary diagnosis was bipolar I. A youth with unipolar depression and a disruptive behavior disorder carried a primary diagnosis of unipolar depression.

schizophrenia, ADHD, and posttraumatic stress disorder (as described in Findling et al., 2001). Youths were referred by local providers or responded to advertisements. Youths and caregivers willing to participate in treatment protocols were included if their initial symptoms appeared to match the enrollment criteria for open trials. Additionally, the sample also included offspring of parents with bipolar disorder who were receiving treatment at an affiliated adult mood disorders clinic.

The community mental health center site consisted of youths and caregivers presenting at a midwestern urban clinic for treatment. Using a consecutive case series design at intake, experimenters asked all youth and caregiver pairs to participate in an assessment research study. All youths—regardless of initial presentation—between the ages of 5 and 18 years were eligible to participate in the current study.

Measures

Schedule for Affective Disorders and Schizophrenia for Children (KSADS). The KSADS is a semistructured interview that queries about symptoms from common Axis I disorders from both the parent and child. The KSADS–PL–Plus amalgamates the mood modules from the Washington University KSADS (Geller et al., 2001) and the KSADS Present and Lifetime Version (Kaufman et al., 1997). The Washington University KSADS includes additional symptoms and associated features of depression and mania beyond those included in the KSADS Present and Lifetime Version. Research assistants were highly trained: Symptom level ratings for new raters were compared with those of a reliable rater for at least five interviews rating along and then five interviews leading. A new rater passed a session if he or she achieved an overall $\kappa \geq .85$ at the item level of the entire interview and a $\kappa = 1.0$ at the diagnostic level. Raters "scored along" with another interviewer on a monthly basis after completing training, and $\kappa \geq .85$ was maintained throughout the project. A new cadre of raters was trained each year, and videotaped interviews were used to avoid drift across years. Research assistants were primarily predoctoral psychology interns or research staff with a master's degree or PhD in psychology or a master's degree in social work. Research assistants conducted assessments at both sites.

Parent-Reported General Behavior Inventory (PGBI). In the PGBI, the original GBI is modified so that all questions in the PGBI query the caregiver about the mood and behavior of his or her offspring (Youngstrom et al., 2001). The PGBI consists of 73 items measuring depressive, hypomanic, and mixed symptoms of mood disorder during the prior year. Participants' answers can range from *never or hardly ever* to *very often or almost constantly* on a 4-point Likert-type scale about their offspring. The Hypomanic/Biphasic (Cronbach's $\alpha = .92$) scale measures symptoms associated with mania in both classical and mixed forms. Present analyses concentrate on the PGBI–10M items.

10-Item Mania General Behavior Inventory. The PGBI–10M was developed from the PGBI using item response theory to determine the 10 best discriminating items from the Hypomanic/Biphasic scale (Youngstrom, Frazier, et al., 2008). Participants' answers can range from *never or hardly ever* to *very often or almost constantly* on a 4-point Likert scale about their offspring's mood symptoms during the prior year (Cronbach's $\alpha = .92$).

Parent Mood Disorder Questionnaire (PMDQ). The PMDQ was developed from the Mood Disorder Questionnaire by changing the target of the items from self to offspring (Wagner et al., 2006). The PMDQ consists of 13 items assessing all of the (hypo)manic symptoms in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; American Psychiatric Association, 2004) using yes-or-no responses, providing a criterion measure of caregiver-reported manic symptoms.

Child Depression Rating Scale–Revised (CDRS–R). The CDRS–R is an adaptation of the Hamilton Rating Scale for Depression designed for use with children and adolescents ranging in age from 5 to 18 years (Poznanski & Mokros, 1996). The CDRS–R consists of 17 items measuring the symptoms of depression. The items are rated between 1 and 7 or 1 and 5, depending on content. Higher scores indicated more severe depression. The CDRS–R was rated by the KSADS interviewer. The CDRS–R is often considered the standard in measuring depressive symptoms in clinical trials for bipolar disorder (Carlson et al., 2003).

Young Mania Rating Scale (YMRS). The YMRS was originally validated in adults (Young, Biggs, Ziegler, & Meyer, 1978). It is now also widely used as a measure of mania symptoms in youths with good evidence that scores have acceptable reliability and construct validity in youths (Fristad, Verducci, Walters, & Young, 2009; Fristad, Weller, & Weller, 1995; Youngstrom, Danielson, Findling, Gracious, & Calabrese, 2002). The YMRS consists of 11 items measuring manic symptoms based on interview of the youth and caregiver by a trained interviewer. The items are rated between 0 to 4 and 0 to 8. Higher scores indicate more severe mania. The YMRS is considered the gold standard for measuring manic symptoms in clinical trials (Carlson et al., 2003).

Child Behavior Checklist. The Achenbach Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) is among the most widely used measures of child and adolescent behavior problems in both research and clinical work. The CBCL consists of 118 items that query about behavior problems in youths between the ages of 6 and 18. Caregivers of youths who were age 5 completed the CBCL 1.5–5.5 Years. The Internalizing score provided a well-established measure of depressive and anxious symptoms.

Procedure

The protocols for EA, EC, and Both groups were similar. Caregivers provided written consent for the youths to participate in the study. Youths provided written assent to participate in the study. The same research assistant interviewed both the caregiver and youth sequentially with the KSADS, CDRS–R, and YMRS. Caregivers completed the PGBI and CBCL as part of an additional battery.

Recruitment for the EC and Both groups occurred during a general clinical intake. During this time, caregivers also completed the PGBI–10M in extracted format. The Both group consisted of individuals who completed both the PGBI–10M, agreed to participate in the assessment study, and presented for the assessment study. The Extracted group received the PGBI–10M as part of standard clinical intake, and de-identified archival data were used for comparison to the other versions.

All cases were reviewed using the Longitudinal Evaluation of All Available Data (LEAD) procedure (Spitzer, 1983). The re-

search assistant met with a licensed clinical psychologist to review the case. During the LEAD meeting, the research assistant presented the KSADS symptoms and diagnoses, family history, and information available from intake (e.g., intake diagnoses, chart review of diagnoses, prior treatment history, and school history). Both the licensed clinical psychologist and the research assistant were blind to the PGBI and the PGBI–10M. Kappas between the KSADS diagnoses and the LEAD diagnoses ranged from .85 (for oppositional defiant disorder) to .93 (for bipolar disorder).

Results

Evaluation of Item Response Theory Assumptions

A confirmatory factor analysis with one latent variable for each of the three samples was fit using Mplus 5.21 (Muthén & Muthén, 1998–2007) to examine whether the items met assumptions of unidimensionality and local independence. The single factor model displayed adequate fit in all three groups (all comparative fit indices $>$ .95 and root-mean-square errors of approximation $<$ .10). Additionally, error correlations between item pairs were all of small magnitude ($<$ .20, following guidelines from Hill et al., 2007; Reeve et al., 2007).

Aim 1: Examination of Sampling Effects: Differential Item Functioning of the 10 Embedded Items Between an Academic Medical Center and a Community Mental Health Center

As expected, the EA and EC groups showed significant and large differences on demographics, socioeconomic status, and clinical characteristics. The academic sample, which was enriched for mood disorders, was more White, had higher socioeconomic status, and had more bipolar I, whereas the community sample had less bipolar I and more spectrum (bipolar II, cyclothymia, and bipolar not otherwise specified). The relative scarcity of bipolar I youths in the EC group and the change in socioeconomic status and ethnicity creates a strong test of the limits on portability across samples. By definition, only bipolar I cases had a history of mania, whereas the rest of the bipolar spectrum could only show at most hypomanic presentations.

For the portability analyses, EA was the reference group, and EC was the focal group. Table 2 displays the item parameters and g^2 goodness-of-fit index for the 10 items. Three items displayed no evidence of DIF. "Rapid mood and energy shifts" (Item 3) and "elated mood or energy with sleep disturbance" (Item 6) were significantly more discriminating in the EA sample, meaning that endorsement of higher categories could occur across a broader spectrum of severity in the EC sample rather than being specific to those with higher levels of mania. "Elated mood only" (Item 2) discriminated significantly better in the EC group than the EA group, meaning that the endorsement of higher response occurred at more distinct severity levels in the EC sample. These four items displayed statistical significant DIF; however, examination of the items' item characteristic curves in Figure 1 indicated that the practical effect size of the difference was minimal.

Items 1, 4, 7, and 8 showed significant differences in the difficulty parameters, ps $<$.05. Caregivers endorsed "mood and energy at the extremes" (Items 7) and "mood switching across days" (Item 8) at lower thresholds in the EC group than the EA, indicating that overall EC caregivers were more likely to endorse higher responses than EA caregivers. EC caregivers endorsed "happiness with energy and hyperactivity" (Item 1) at significantly lower thresholds than the EA caregivers. "Happiness with energy" (Item 4) was significantly more difficult for EC group than EA group at the extreme scores. The differences in thresholds were typically in the small effect size range.

Figure 2 shows that even though many of the items displayed differential functioning between the two settings, as a scale the 10 items were functioning similarly across settings. Small item-level differences in opposite directions canceled out at the scale total level. In both samples, the 10 items produced nearly identical observed scores for individuals with the same severity of mania.

Aim 2: Examination of Item Context Effects: Differential Item Functioning of the 10 Items Embedded Versus Extracted

For the context effect analyses, EC was the reference group and Extracted was the focal group. Table 3 displays the item parameters and g^2 goodness-of-fit for the 10 items. After the false discovery rate was controlled, only "mood and energy always at the extremes" (Item 7) yielded lower scores for the extracted items compared with when the items were embedded. Caregivers responded to Item 7 at lower thresholds when it was not in the context of 63 other mood items. A single item with small threshold differences did not substantially alter the 10 items' functioning together as a scale. Therefore, context effects did not appear to substantively change the overall performance of the 10 items when administered in an extracted form.

Aim 3: Examination of Construct Validity of the 10-Item GBI

Convergent validity. Table 4 shows Pearson correlation coefficients using the Both group because they received all measures, including both the PGBI–10M and the 10 items embedded in the PGBI. There was a significant positive, strong correlation between the PGBI–10M and PGBI versions of the 10 items, $r = .64$, $p < .05$. These administrations were separated by approximately 1 week (median = 8 days). The PGBI–10M and the PGBI had significant, strong, positive correlations with the YMRS (rs = .46 and .49, $p < .05$), and PMDQ (rs = .48 and .74, $p < .05$), consistent with showing convergent validity for mania.

Discriminant validity. The PGBI–10M demonstrated large correlations with other established parent-reported measures of mania (e.g., PMDQ) and with interview ratings of manic symptoms that were made blind to the PGBI–10M scores (see Table 4). The PGBI–10M also showed significant correlations with the measures of depressed mood and internalizing, as expected given that (a) many of the items on the PGBI–10M are from the "biphasic/mixed" component of the original GBI (Depue, 1981) and (b) youths with bipolar disorder showed elevated depressed as well as manic symptoms. Even so, the PGBI–10M showed significantly higher correlations with the YMRS than the CDRS–R (both based on interviewer ratings blind to the questionnaire scores), $t(156) = 2.19$, $p < .05$. for the extracted, and $t = 1.07$, *ns*, for the embedded. Similarly, PGBI–10M score correlations were lower with the

Table 2
Discrimination and Difficulty Parameter Estimates From Differential Item Functioning Results Comparing Embedded Academic to Embedded Community Samples

Item & Content/Group	Discrimination	Threshold 1	Threshold 2	Threshold 3	Discrimination DIF χ^2 (p)	Threshold DIF χ^2 (p)
No differential item functioning						
Item 5: Happy + Energy + Rage						
EA	1.93	0.17	1.11	1.79	.0 (1.00)	8.6 (.04)
EC	1.92	0.32	1.20	2.10		
Item 9: Happy + Energy + Anger						
EA	2.12	−0.20	0.99	1.77	2.5 (.11)	2.8 (.42)
EC	2.49	−0.11	0.89	1.74		
Item 10: Racing thoughts						
EA	1.64	0.48	1.78	2.69	3.8 (.05)	7.4 (.06)
EC	1.28	0.33	1.68	2.85		
More discriminating in EA than EC						
Item 3:Rapid mood/energy shift						
EA	2.17	−0.53	0.57	1.40	7.5 (.01)*	7.5 (.06)
EC	1.63	−0.75	0.39	1.42		
Item 6: Happiness/Energy + Sleep disturbance						
EA	2.17	0.19	1.21	1.85	3.9 (.05)*	11.3 (.01)
EC	1.74	0.07	1.01	1.88		
Less discriminating in EA than EC						
Item 2: Happy						
EA	1.58	0.65	1.75	2.76	4.2 (.04)*	7.9 (.05)
EC	2.02	0.69	1.85	2.72		
More difficult in EA than EC						
Item 7: Mood + Energy at extremes	1.87	−0.48	0.72	1.55	1.6 (.21)	39.2 (<.01)*
EA	1.64	−0.80	0.21	1.24		
EC						
Item 8: Mood switching across days						
EA	2.25	0.03	1.26	1.94	2.3 (.13)	15.8 (<.01)*
EC	2.65	0.27	1.11	1.79		
More difficult in EC than EA						
Item 1: Happy + Energy + Hyperactivity						
EA	1.62	−0.20	0.69	1.36	5.0 (.03)*	27.9 (<.01)
EC	2.06	−0.02	0.81	1.73		
Item more difficult at average and extremely high levels in EC, but more difficult at high levels in EA						
Item 4: Happy + Energy						
EA	2.13	0.30	1.37	2.03	3.2 (.07)	11.5 (.01)*
EC	2.61	0.44	1.28	2.17		

* Indicates significantly different after Benjamini–Hochberg correction. EA = Embedded Academic; EC = Embedded Community.

CBCL Internalizing than with the other parent-reported mania scale, $t(156) = 4.25$, $p < .00005$, for the embedded, and $t = 0.38$, *ns*, for the extracted scale. Multiple linear regressions indicated that the partial correlation among the PGBI–10M, and each of these measures was substantially reduced after controlling for the number of comorbid diagnoses, ADHD, and disruptive behavior disorders, ps $< .05$. The correlations between these scales suggest that scores on the PGBI–10M were also being influenced by a youth's current depressed mood state and commonly overlapping comorbid diagnoses at study entry.

Aim 4: Examination of the Diagnostic Efficiency of the 10-Item GBI

Receiver operating characteristic (ROC) curves examined whether the PGBI–10M could distinguish between youths with PBD and all other youths. ROC compares the sensitivity and false alarm rate (1−specificity), which can best be quantified by the area under the ROC curve (AUROC; Altman & Bland, 1994). An AUROC of .50 would indicate chance performance or an inability to distinguish youths with PBD from other youths. These analyses examine the discriminative validity of the PGBI-10M. The sample design—with high rates of comorbidity, high rates of diagnoses likely to generate false positive responses, and relatively low rates of extreme mania—creates a conservative but clinically realistic scenario for evaluating this aspect of validity. The PGBI–10M discriminated PBD from all other diagnoses significantly better than chance, AUROC = .79, $p < .05$, 95% confidence interval (CI) [.69, .90]. The 10 items embedded in the PGBI also detected PBD significantly better than chance, AUROC = .80, $p < .05$, 95% CI [.71, .89]. The PMDQ detected PBD significantly better than chance, AUROC = .84, $p < .05$, 95% CI [.75, .92]. The ability of the PGBI–10M to discern PBD was not significantly different than the 10 items embedded in the PGBI, $z = .06$, $p = .95$, or the PMDQ, $z = -.71$, $p = .48$ (using the test from Hanley & McNeil, 1983). The PGBI–10M distinguished PBD compared with depression, AUROC = .78, $p < .01$, 95% CI [.65, .92], and

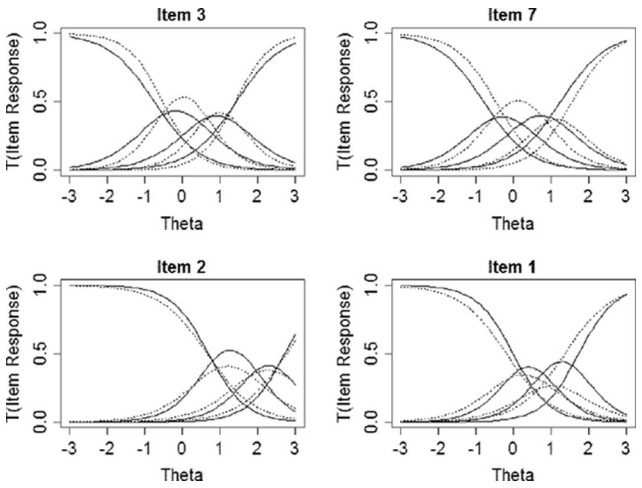

Figure 1. Boundary response functions for selected items showing Differential item functioning between the Embedded Academic (EA) and Embedded Community (EC) samples. Solid line is EC sample. Dotted line is EA sample. Item 3 is more discriminating in EA than EC. Item 7 is more difficult in EA than EC. Item 2 is less discriminating in EA than EC. Item 1 is less difficult in EA than EC.

any disruptive behavior disorder, AUROC = .78, $p < .01$, 95% CI [.65, .92].

Discussion

The first specific aim of this project was to examine the portability of the 10 best discriminating PGBI items moving between an academic medical center and community mental health center. These sites differed markedly in terms of demographic features of their samples such as socioeconomic status and caregiver educational level, as well as in terms of clinical characteristics such as the rate of bipolar disorder or the proportion of spectrum bipolar diagnoses versus bipolar I diagnosis. Despite these differences in sample composition, data indicated that the 10 items function similarly across samples and context. When comparing performance in a sample in which individuals have a higher income and are primarily White, and in which probands often have been selected for mood disorder versus a sample with lower income, is primarily African American, and has lower rates of mood disorder, we found that the 10 items showed little evidence of DIF and nonsignificant differences in total score functioning. At the item level, querying "rapid mood/energy shifts" and "elated mood with sleep disturbance" was mildly less discriminating in the community mental health sample. Querying "elated mood only" was slightly more discriminating in the community mental health sample. Caregivers were more likely to endorse "mood and energy at the extremes" in the community mental health sample than at the academic medical center, while they were less likely to endorse "elated mood with hyperactivity and high energy" at the community mental health center. Visual examination of the statistically significant effects suggested that differences in item functioning were small. Additionally, the item-level differences appeared to balance themselves across the scale. After controlling for mean differences, the total observed score represented equivalent severities of mania between the two samples, even though individual items showed differences across the two samples.

The second specific aim was to examine whether context effects occur when extracting the 10 items from the full parent-reported GBI. The findings indicated that context did not have a strong effect on caregiver responses to the 10 items. Nine of the 10 items showed no significant differences in their relationship to mania or to the amount of mania required to endorse any particular response when they were administered by themselves or within the context of the full-length PGBI. The one exception was the item querying "extreme mood and energy." In the extracted, free-standing 10 items, caregivers were slightly more likely to endorse higher response categories at similar levels of mania. These results appear consistent with the suggestion by Steinberg (2001) that precise items are less likely to be affected by context. Item response is most likely due to respondents pooling prior memories, evaluating the consistency of those memories, and evaluating the similarities amongst the memories (Tourangeau, Rips, & Rasinski, 2000). Vague items are more likely to pull for memories that are not consistent or similar. The detail of the GBI items probably reduces the role of context for most items.

The third and fourth specific aims were to examine the validity of the extracted PGBI–10M scale scores. The results indicate the PGBI–10M scores are measuring the construct of mania in youths based on the high agreement with clinician ratings of manic symptoms and caregiver reported mania on another rating scale. However, the PGBI–10M scores overlapped with depressive symptomatology, as well as comorbid disorders such as ADHD and disruptive behavior disorders. Despite these correlations, the clinically typical rates of comorbidity (Kowatch et al., 2005), and the high rates of diagnoses often challenging to differentiate from bipolar disorder (Kim & Miklowitz, 2002), the PGBI–10M was able to identify youths with PBD significantly better than chance from all other youths presenting to the clinics. More focal comparisons demonstrated that the PGBI–10M also could discriminate bipolar from unipolar depression or ADHD.

Strengths of this study include the large, multisite, diverse sample of youths with reports of mania symptoms, and examina-

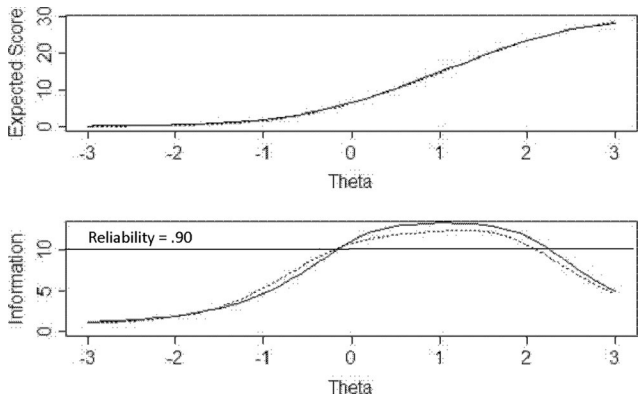

Figure 2. Test characteristic and test information curves comparing the 10 mania items from Parent-Reported General Behavior Inventory (PGBI–10M) of the Embedded Academic (dotted line) sample to the same 10 items in the Embedded Community (solid line) sample.

Table 3
Differential Item Functioning Results Comparing Embedded Community to the Extracted Community Samples

Item & Content/Group	Discrimination	Threshold 1	Threshold 2	Threshold 3	Discrimination DIF χ^2 (p)	Threshold DIF χ^2 (p)
No differential item functioning						
Item 1: Happy + Energy + Hyperactivity						
Embedded	1.58	−0.20	−0.78	1.78	0.0 (1.00)	13.3 (<.01)
Extracted	1.59	−0.20	0.71	1.38		
Item 2: Happy						
Embedded	1.80	0.44	1.54	2.54	1.8 (.18)	4.9 (.18)
Extracted	1.53	0.66	1.80	2.84		
Item 3:Rapid mood/Energy shift						
Embedded	2.09	−0.49	0.71	1.56	0.1 (.75)	5 (.17)
Extracted	2.15	−0.56	0.55	1.40		
Item 4: Happy + Energy						
Embedded	2.42	0.35	1.23	2.21	2.0 (.16)	9.2 (.03)
Extracted	2.07	0.30	1.40	2.07		
Item 5: Happy + Energy + Rage						
Embedded	2.45	0.00	0.95	1.72	5.2 (.02)	9.4 (.02)
Extracted	1.91	0.18	1.14	1.82		
Item 6: Happiness/Energy + Sleep disturbance						
Embedded	1.89	0.23	1.16	2.04	1.1 (.29)	4.2 (.24)
Extracted	2.12	0.18	1.22	1.88		
Item 8: Mood switching across days						
Embedded	2.21	−0.02	1.11	2.03	0.0 (1.00)	5.5 (.13)
Extracted	2.24	0.03	1.27	1.95		
Item 9: Happy + Energy + Anger						
Embedded	2.30	−0.27	0.85	1.78	0.5 (.48)	4 (.26)
Extracted	2.14	−0.20	0.99	1.77		
Item 10: Racing thoughts						
Embedded	1.57	0.26	1.45	2.45	0.0 (1.00)	12.3 (<.01)
Extracted	1.58	0.49	1.82	2.76		
More mania to endorse higher responses if item is embedded						
Item 7: Mood + Energy at extremes						
Embedded	1.86	−0.29	1.04	2.06	0.0 (1.00)	29 (<.01)*
Extracted	1.88	−0.52	0.68	1.51		

* Indicates significantly different after Benjamini–Hochberg correction.

tion of one of the best performing instruments currently available (Youngstrom et al., 2009). In the present study, the PGBI–10M performed well across sites, suggesting that it is portable and resistant to context effects. Additionally, the current study reflects one of the first attempts to study item-level functioning in youths with PBD.

The diverse sample is also a limitation. Due to the differences in socioeconomic status, ethnicity, and diagnostic differences be-

Table 4
Correlation Matrix of the Embedded and Extracted 10 Items for the Both Group With Criterion Measures (N = 159)

Variable	PGBI–10M extracted	PGBI embedded	t test of dependent correlations[a] (df = 156)
PGBI embedded	.64***	—	
Clinician-administered (blind to rating scale scores)			
YMRS (Mania)	.46***	.49***	t = 0.52
CDRS–R (Depression)	.29**	.41***	t = 1.93
	t = 2.19*	t = 1.07	
Caregiver rating scales			
Mood Disorder Questionnaire (PMDQ Mania)	.48***	.74***	t = 5.54***
CBCL Internalizing	.45***	.47***	t = 0.34
	t = 0.38	t = 4.25***	

Note. PGBI = Parent-Reported General Behavior Inventory; PGBI–10M = 10 mania items from PGBI; YMRS = Young Mania Rating Scale; CDRS–R = Child Depression Rating Scale–Revised; PMDQ = Parent Mood Disorder Questionnaire; CBCL = Child Behavior Checklist.
[a] The t test compares whether the correlation is significantly different for the embedded versus extracted versions given the same criterion variable.
* $p < .05$. ** $p < .005$. *** $p < .0005$.

tween the academic medical center and the community mental health center, the item-level differences cannot be attributed with certainty to any single factor. Although the effect sizes are small, the sample differences prevent a conclusion about whether certain items (e.g., elated mood) are better predictors for Whites or African Americans, or for lower versus higher socioeconomic status, or whether differences are due to differences in respondents' reading ability. Item response theory allows for group differences in mean scores due to differences in diagnostic discrepancies and comorbidity patterns, because it places items and individuals at equivalent trait levels prior to examining DIF (Thissen, Steinberg, & Gerrard, 1986). Items that evaluate straightforward and easily observable behavior might be less susceptible to context and sampling effects than vague items (e.g., Steinberg, 2001).

Future studies should examine what the item-level differences are due to, such as differences in race/ethnicity, differences in socioeconomic status, or potentially differences in reading level. Knowing these differences and whether they have large effect sizes could aid clinicians in determining lines of questioning and the weight to place on different symptoms conditioned upon easily identifiable demographic information. Additionally, examining reasonable cut scores and developing diagnostic likelihood ratios (e.g., Straus et al., 2005) could aid in clinical prediction of PBD. Ideally, these will be based on large enough samples to provide good estimates of optimal thresholds and small standard errors and to define multiple thresholds to preserve more information from the raw scores. Replication in other clinical settings with different levels of severity of bipolar presentation, such as inpatient units or public schools, would be important to understand if the items continue to behave similarly even at the extremes of the latent factor of mania. Finally, it is worth noting that the PGBI–10M concentrates on manic and mixed symptoms, which are only a small—albeit more diagnostically specific—aspect of bipolar disorder. A comprehensive approach to the assessment of bipolar disorder would also include scales pertaining to depression, anxiety, and perhaps quality of life or other domains of functioning relevant to case formulation and evaluation of outcomes.

Even so, the present analyses do much to enhance confidence that the PGBI–10M performs in a robust manner even when the items are used in the brief, extracted format and even when employed in diverse settings such as urban community mental health centers. Results indicate that the brief version of the scale continues to provide clinically useful information in the assessment of pediatric bipolar disorder across a broad range of clinical settings.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles.* Burlington: University of Vermont Research Center for Children, Youth, and Families.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 3: Receiver operating characteristic plots [Statistics note]. *British Medical Journal, 309,* 188.

American Psychiatric Association. (2004). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

American Psychiatric Association. (2011). *Diagnostic and statistical manual of mental disorders* (5th ed.). Retrieved March 17, 2011, from www.dsm5.org

Birmaher, B., Axelson, D., Strober, M., Gill, M. K., Valeri, S., Chiappetta, L., . . . Keller, M. (2006). Clinical course of children and adolescents with bipolar spectrum disorders. *Archives of General Psychiatry, 63,* 175–183. doi:10.1001/archpsyc.63.2.175

Blader, J. C., & Carlson, G. A. (2007). Increased rates of bipolar disorder diagnoses among U.S. child, adolescent, and adult inpatients, 1996–2004. *Biological Psychiatry, 62,* 107–114. doi:10.1016/j.biopsych.2006.11.006

Bowring, M. A., & Kovacs, M. (1992). Difficulties in diagnosing manic disorders among children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 31,* 611–614. doi:10.1097/00004583-199207000-00006

Carlson, G. A., Jensen, P. S., Findling, R. L., Meyer, R. E., Calabrese, J., DelBello, M. P., . . . Youngstrom, E. A. (2003). Methodological issues and controversies in clinical trials with child and adolescent patients with bipolar disorder: Report of a consensus conference. *Journal of Child and Adolescent Psychopharmacology, 13,* 13–27. doi:10.1089/104454603321666162

Depue, R. A. (1981). A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: A conceptual framework and five validation studies. *Journal of Abnormal Psychology, 90,* 381–437. doi:10.1037/0021-843X.90.5.381

Findling, R. L., Gracious, B. L., McNamara, N. K., Youngstrom, E. A., Demeter, C. A., Branicky, L. A., & Calabrese, J. R. (2001). Rapid, continuous cycling and psychiatric co-morbidity in pediatric bipolar I disorder. *Bipolar Disorders, 3,* 202–210. doi:10.1034/j.1399-5618.2001.30405.x

Fristad, M. A., Verducci, J. S., Walters, K., & Young, M. E. (2009). Impact of multifamily psychoeducational psychotherapy in treating children aged 8 to 12 years with mood disorders. *Archives of General Psychiatry, 66,* 1013–1020. doi:10.1001/archgenpsychiatry.2009.112

Fristad, M. A., Weller, R. A., & Weller, E. B. (1995). The Mania Rating Scale (MRS): Further reliability and validity studies with children. *Annals of Clinical Psychiatry, 7,* 127–132. doi:10.3109/10401239509149039

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

Geller, B., Zimerman, B., Williams, M., DelBello, M. P., Frazier, J., & Beringer, L. (2002). Phenomenology of prepubertal and early adolescent bipolar disorder: Examples of elated mood, grandiose behaviors, decreased need for sleep, racing thoughts and hypersexuality. *Journal of Child and Adolescent Psychopharmacology, 12,* 3–9. doi:10.1089/10445460252943524

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148,* 839–843. doi:10.1016/S1076-6332(97)80161-4

Henry, D. B., Pavuluri, M. N., Youngstrom, E., & Birmaher, B. (2008). Accuracy of brief and full forms of the Child Mania Rating Scale. *Journal of Clinical Psychology, 64,* 368–381. doi:10.1002/jclp.20464

Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., Burwinkle, T. M., & Varni, J. W. (2007). Practical issues in the application of item response theory: A demonstration using items from the Pediatric Quality of Life Inventory (PedsQL) 4.0 generic core scales. *Medical Care, 45*(5, Suppl 1), S39–S47. doi:10.1097/01.mlr.0000259879.05499.eb

Hirschfeld, R. M., Holzer, C., Calabrese, J. R., Weissman, M., Reed, M., Davies, M., . . . Hazard, E. (2003). Validity of the Mood Disorder Questionnaire: A general population study. *The American Journal of Psychiatry, 160,* 178–180. doi:10.1176/appi.ajp.160.1.178

Hirschfeld, R. M., Williams, J. B., Spitzer, R. L., Calabrese, J. R., Flynn, L., Keck, P. E., Jr., . . . Zajecka, J. (2000). Development and validation of a screening instrument for bipolar spectrum disorder: The Mood

Disorder Questionnaire. *The American Journal of Psychiatry, 157,* 1873–1875. doi:10.1176/appi.ajp.157.11.1873

Jenkins, M. M., Youngstrom, E. A., Washburn, J. J., & Youngstrom, J. K. (in press). Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice, 42,* 121–129. doi:10.1037/a0022506

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

Kim, E. Y., & Miklowitz, D. J. (2002). Childhood mania, attention-deficit hyperactivity disorder and conduct disorder: A critical review of diagnostic dilemmas. *Bipolar Disorders, 4,* 215–225. doi:10.1034/j.1399-5618.2002.01191.x

Kowatch, R. A., Youngstrom, E. A., Danielyan, A., & Findling, R. L. (2005). Review and meta-analysis of the phenomenology and clinical characteristics of mania in children and adolescents. *Bipolar Disorders, 7,* 483–496. doi:10.1111/j.1399-5618.2005.00261.x

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines.* Newbury Park, CA: Sage.

Kraepelin, E. (1976). *Manic-depressive insanity and paranoia* (R M. Barclay, Trans.). New York: Arno Press. (Original work published 1921)

Lewinsohn, P. M., Klein, D. N., & Seeley, J. R. (1995). Bipolar disorders in a community sample of older adolescents: Prevalence, phenomenology, comorbidity, and course. *Journal of the American Academy of Child & Adolescent Psychiatry, 34,* 454–463. doi:10.1097/00004583-199504000-00012

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Merikangas, K. R., He, J. P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., . . . Swendsen, J. (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS–A). *Journal of the American Academy of Child and Adolescent Psychiatry, 49,* 980–989. doi:10.1016/j.jaac.2010.05.017

Merikangas, K. R., Jin, R., He, J. P., Kessler, R. C., Lee, S., Sampson, N. A., . . . Zarkov, Z. (2011). Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Archives of General Psychiatry, 68,* 241–251. doi:10.1001/archgenpsychiatry.2011.12

Merikangas, K. R., & Pato, M. (2009). Recent developments in the epidemiology of bipolar disorder in adults and children: Magnitude, correlates, and future directions. *Clinical Psychology: Science and Practice, 16,* 121–133. doi:10.1111/j.1468-2850.2009.01152.x

Miller, C. J., Johnson, S. L., & Eisner, L. (2009). Assessment tools for adult bipolar disorder. *Clinical Psychology: Science and Practice, 16,* 188–201. doi:10.1111/j.1468-2850.2009.01158.x

Miller, C. J., Klugman, J., Berv, D. A., Rosenquist, K. J., & Ghaemi, S. N. (2004). Sensitivity and specificity of the Mood Disorder Questionnaire for detecting bipolar disorder. *Journal of Affective Disorders, 81,* 167–171. doi:10.1016/S0165-0327(03)00156-3

Moreno, C., Laje, G., Blanco, C., Jiang, H., Schmidt, A. B., & Olfson, M. (2007). National trends in the outpatient diagnosis and treatment of bipolar disorder in youth. *Archives of General Psychiatry, 64,* 1032–1039. doi:10.1001/archpsyc.64.9.1032

Muthén, L. K., & Muthén, B. O. (1998–2007). Mplus user's guide (Version 5.2). Los Angeles, CA: Authors.

Neighbors, H. W., Jackson, J., Campbell, L., & Williams, D. (1989). The influence of racial factors on psychiatric diagnosis: A review and suggestions for research. *Community Mental Health Journal, 25,* 301–311. doi:10.1007/BF00755677

Pavuluri, M. N., Henry, D. B., Devineni, B., Carbray, J. A., & Birmaher, B. (2006). Child Mania Rating Scale: Development, reliability, and validity. *Journal of the American Academy of Child & Adolescent Psychiatry, 45,* 550–560. doi:10.1097/01.chi.0000205700.40700.50

Pliszka, S. R., Sherman, J. O., Barrow, M. V., & Irick, S. (2000). Affective disorder in juvenile offenders: A preliminary study. *The American Journal of Psychiatry, 157,* 130–132.

Poznanski, E. O., & Mokros, H. B. (1996). *Children's Depression Rating Scale-Revised (CDRS-R).* Los Angeles, CA: Western Psychological Services.

Quinn, C. A., & Fristad, M. A. (2004). Defining and identifying early onset bipolar spectrum disorder. *Current psychiatry reports, 6*(2), 101–107. doi:10.1007/s11920-004-0049-1

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D., 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

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18,* 169–184. doi:10.1002/mpr.289

Schuman, H., Presser, S., & Ludwig, J. (1981). Context effects on survey responses to questions about abortion. *Public Opinion Quarterly, 45,* 216–223. doi:10.1086/268652

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24,* 399–411. doi:10.1016/0010-440X(83)90032-9

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81,* 332–342. doi:10.1037/0022-3514.81.2.332

Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). New York, NY: Churchill Livingstone.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128. doi:10.1037/0033-2909.99.1.118

Tillman, R., & Geller, B. (2005). A brief screening tool for a prepubertal and early adolescent bipolar disorder phenotype. *The American Journal of Psychiatry, 162,* 1214–1216. doi:10.1176/appi.ajp.162.6.1214

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York, NY: Cambridge University Press.

Van Meter, A. R., Moreira, A. L., & Youngstrom, E. A. (2011). Meta-analysis of epidemiologic studies of pediatric bipolar disorder. *The Journal of Clinical Psychiatry.* doi:10.4088/JCP.10m06290

Wagner, K. D., Hirschfeld, R. M. A., Emslie, G. J., Findling, R. L., Gracious, B. L., & Reed, M. L. (2006). Validation of the Mood Disorder Questionnaire for bipolar disorders in adolescents. *Journal of Clinical Psychiatry, 67,* 827–830. doi:10.4088/JCP.v67n0518

Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: Reliability, validity, and sensitivity. *British Journal of Psychiatry, 133,* 429–435. doi:10.1192/bjp.133.5.429

Youngstrom, E. A., Birmaher, B., & Findling, R. L. (2008). Pediatric bipolar disorder: Validity, phenomenology, and recommendations for diagnosis. *Bipolar Disorders, 10*(Suppl 1, Pt 2), 194–214. doi:10.1111/j.1399-5618.2007.00563.x

Youngstrom, E. A., Danielson, C. K., Findling, R. L., Gracious, B. L., & Calabrese, J. R. (2002). Factor structure of the Young Mania Rating Scale for use with youths ages 5 to 17 years. *Journal of Clinical Child and Adolescent Psychology, 31,* 567–572. doi:10.1207/153744202320802232

Youngstrom, E. A., Findling, R. L., Danielson, C. K., & Calabrese, J. R.

(2001). Discriminative validity of parent report of hypomanic and depressive symptoms on the General Behavior Inventory. *Psychological Assessment, 13,* 267–276. doi:10.1037/1040-3590.13.2.267

Youngstrom, E. A., Frazier, T. W., Demeter, C., Calabrese, J. R., & Findling, R. L. (2008). Developing a 10-item mania scale from the Parent General Behavior Inventory for Children and Adolescents. *Journal of Clinical Psychiatry, 69,* 831–839. doi:10.4088/JCP.v69n0517

Youngstrom, E. A., Freeman, A. J., & Jenkins, M. M. (2009). The assessment of children and adolescents with bipolar disorder. *Child and Adolescent Psychiatric Clinics of North America, 18,* 353–390. doi:10.1016/j.chc.2008.12.002

Youngstrom, E. A., & Green, K. W. (2003). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. *Educational and Psychological Measurement, 63,* 279–295. doi:10.1177/0013164403253226

Youngstrom, E. A., Mash, E. J., & Barkley, R. A. (2007). *Pediatric bipolar disorder Assessment of childhood disorders* (4th ed., pp. 253–304). New York, NY: Guilford Press.

Youngstrom, E. A., Meyers, O., Demeter, C. A., 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*(6), 507–517. doi:10.1111/j.1399-5618.2005.00269.x

Youngstrom, E. A., Meyers, O., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006). Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry, 60,* 1013–1019. doi:10.1016/j.biopsych.2006.06.023

Youngstrom, E. A., Youngstrom, J. K., & Starr, M. (2005). Bipolar diagnoses in community mental health: Achenbach Child Behavior Checklist profiles and patterns of comorbidity. *Biological Psychiatry, 58,* 569–575. doi:10.1016/j.biopsych.2005.04.004

Received September 17, 2010
Revision received August 15, 2011
Accepted August 17, 2011 ■