tapraid5/arc-arc/arc-arc/arc00115/arc0025d15z | xppws | S=1 | 10/26/15 | Art: 2015-1063

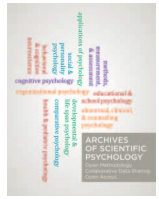
Archives of Scientific Psychology 2015, 3, 000
DOI: http://dx.doi.org/10.1037/arc0000024

2169-3269

American Psychological Association

Archives of Scientific Psychology

www.apa.org/pubs/journals/arc



Multivariate Meta-Analysis of the Discriminative Validity of Caregiver, Youth, and Teacher Rating Scales for Pediatric Bipolar Disorder: Mother Knows Best About Mania

AQ: au
AQ: 1

Eric A. Youngstrom and Jacquelynne Genzlinger
University of North Carolina at Chapel Hill

Gregory Egerton
University at Buffalo, The State University of New York

Anna R. Van Meter
Yeshiva University

A B S T R A C T

The past 2 decades have seen a rapid increase in the amount of research on bipolar disorder in children and adolescents, including studies that look at the accuracy of symptom checklists as a way of telling if a youth might have bipolar disorder. How accurate are these checklists? Does accuracy change if they are completed by the youth or a teacher instead of the primary caregiver? Are checklists that focus specifically on symptoms of mania more accurate than checklists with more general content—typical of older measures? How much does the performance of checklists change depending on whether the sample only includes youths seeking treatment, versus including a healthy comparison group? We addressed these research questions by systematically reviewing major publication databases (PsycINFO, PubMed, and GoogleScholar) and looking at 4,094 hits based on our search. We looked for studies that reported enough information to (a) estimate the size of the difference in checklist scores ("effect size") between cases with versus without research diagnoses of bipolar disorder for, (b) youths 18 years of age or younger, and (c) including at least 10 cases with bipolar disorder. Because we wanted to compare caregiver, teacher, and youth report on the same measures, we used a newer statistical technique, multivariate meta-analysis, to combine and compare results within as well as across studies. We found 63 effect sizes from 8 checklists used in 27 separate samples, including 11,941 youths, of whom 1,834 had diagnoses of bipolar disorder. Overall, checklists did a good job separating cases with bipolar from other youths, with an effect size of 1.05, meaning that bipolar cases scored more than a *SD* higher. Caregiver report was the most accurate across all checklists, performing significantly better than youth or teacher report. Scales focusing on manic symptoms also outperformed general symptom checklists. Sample composition also changed the accuracy of the checklists a great deal: Many studies either included healthy children or excluded youths with diagnoses that are difficult to tell apart from bipolar disorder. These studies gave an overly optimistic sense of how well the checklist might do at identifying youths with bipolar disorder in most clinical settings. Three checklists have shown validity in multiple studies and appear accurate enough to be helpful in improving diagnosis in clinical practice.

Eric A. Youngstrom and Jacquelynne E. Genzlinger, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill; Gregory A. Egerton, Department of Psychology, University at Buffalo, The State University of New York; Anna R. Van Meter, Ferkauf Graduate School, Yeshiva University.

This work was supported in part by a grant from the Lindquist Foundation. We thank Camille Sowder, Mian-Li Ong, Karen Bourne, and Ericka McKinney for their help with coding and contributions to preliminary analyses. We thank Angela Bardeen, Ph.D., for consultation with the search strategy and tracking of coding. Thanks also to Wolfgang Viechtbauer, Ph.D., for consultation around the *Metafor* software package and technical aspects of multivariate meta-analysis.

The authors have made available for use by others the data that underlie the analyses presented in this article (see Youngstrom, 2014), thus, allowing replication and potential extensions of this work by qualified researchers. Next users are obligated to involve the data originators in their publication plans, if the originators so desire.

For further discussion on this topic, please visit the *Archives of Scientific Psychology* online public forum at http://arcblog.apa.org.

Correspondence concerning this article should be addressed to Eric A. Youngstrom, Department of Psychology, University of North Carolina at Chapel Hill, CB #3270, Davie Hall, Chapel Hill, NC 27599-3270. E-mail: eay@unc.edu

SCIENTIFIC ABSTRACT

AQ: 2

To meta-analyze the diagnostic efficiency of checklists for discriminating pediatric bipolar disorder (PBD) from other conditions. Hypothesized moderators included (a) informant—we predicted caregiver report would produce larger effects than youth or teacher report; (b) scale content—scales that include manic symptoms should be more discriminating; and (c) sample design—samples that include healthy control cases or impose stringent exclusion criteria are likely to produce inflated effect sizes. Searches in PsycINFO, PubMed, and GoogleScholar generated 4,094 hits. Inclusion criteria were (a) sufficient statistics to estimate a standardized effect size, (b) age 18 years or less, and (c) at least 10 cases (d) with diagnoses of PBD based on semistructured diagnostic interview. Multivariate mixed regression models accounted for nesting of multiple effect sizes from different informants or scales within the same sample. Data included 63 effect sizes from 8 rating scales across 27 separate samples (N = 11,941 youths, 1,834 with PBD). The average effect size was g = 1.05. Random effect variance components within study and between study were significant, $ps < .00005$. Informant, scale content, and sample design all explained significant unique variance, even after controlling for design and reporting quality. Checklists have clinical utility for assessing PBD. Caregiver reports discriminated PBD significantly better than teacher and youth self report, although all 3 showed discriminative validity. Studies using "distilled" designs with healthy control comparison groups, or stringent exclusion criteria, produced significantly larger effect size estimates that could lead to inflated false positive rates if used as described in clinical practice.

Keywords: bipolar disorder, children and adolescents, sensitivity and specificity, meta-analysis, mania
Supplemental materials: http://dx.doi.org/10.1037/arc0000024.supp
Data repository: http://dx.doi.org/10.3886/ICPSR36245.v1

(Youngstrom, 2015)

The diagnosis of bipolar disorder in children and adolescents has been one of the most contentious issues in child mental health over the past two decades (Carlson & Klein, 2014; Biederman, Klein, Pine, & Klein, 1998). The questions of whether bipolar disorder could manifest before puberty, whether the same criteria should be used for children as for adults, and the validity and importance of collateral reports about mood and behavior by caregivers and teachers have guided a growing body of research (Fristad & Macpherson, 2014; Geller & Luby, 1997; Youngstrom, Birmaher, & Findling, 2008). Given these longstanding debates and the growing literature on the topic of pediatric bipolar disorder, it is opportune to undertake a quantitative review of the literature on assessment of pediatric bipolar disorder. A meta-analysis also can address larger themes of cross-informant validity and fundamental research design issues that cut across the wider domains of clinical assessment.

Importance of Accurate Identification of Bipolar Disorder

A substantial portion of mood disorders fall along the spectrum of bipolar disorders, which includes not only bipolar I, but also bipolar II, cyclothymic disorder, and bipolar not otherwise specified (now "other specified bipolar and related disorders;" American Psychiatric Association, 2013). Both longitudinal and epidemiological studies indicate that at least a third of serious mood disorders follow a bipolar course (Angst et al., 2011, 2012; Merikangas et al., 2007). Bipolar disorder also needs different treatment strategies (Yatham et al., 2005). It is not just the difference between bipolar and unipolar depression that matters for prognosis or treatment prescription. Disruptive behavior disorders—oppositional defiant disorder, conduct disorder, and the new diagnosis of dysregulated mood disorder with dysphoria—also are challenging to distinguish from bipolar disorder (Axelson et al., 2012), and they would indicate substantially different approaches to treatment (cf. Eyberg, Nelson, & Boggs, 2008; Fristad & Macpherson, 2014). The same is true of attention-deficit/hyperactivity disorder (ADHD), which has been the nexus of extensive debate both because of overlapping symptoms as well as concerns about possible iatrogenic effects of using stimulants when the person has bipolar disorder (Carlson, 2003; Scheffer, Kowatch, Carmody, & Rush, 2005; Youngstrom, Arnold, & Frazier, 2010).

Despite the risks associated with misdiagnosis, clinical practice often does an exceptionally poor job of recognizing bipolar disorder. A meta-analysis comparing clinical diagnoses of children and adolescents to structured or semistructured diagnoses found an average κ of .27 (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). Dismayingly, the κ was even lower for bipolar disorder, K = .08. Similarly, comparisons of the accuracy of clinical diagnoses versus research consensus diagnoses—arguably even more valid than semistructured interviews alone (Spitzer, 1983)—have found that bipolar diagnoses are among the least accurate (Jensen-Doss, Youngstrom, Youngstrom, Feeny, & Findling, 2014), particularly among ethnic minority groups (e.g., Delbello, Lopez-Larson, Soutullo, & Strakowski, 2001). Misdiagnosis reduces the likelihood of appropriate intervention (McClellan, Kowatch, & Findling, 2007).

Potential Role of Rating Scales for Discriminating Between Bipolar and Other Diagnoses

Rating scales and checklists can potentially improve diagnosis, providing tools that are inexpensive and depend less on training to be implemented consistently across settings (cf. Drotar, Stein, & Perrin, 1995; Jenkins, Youngstrom, Washburn, & Youngstrom, 2011), and often have good psychometric properties within the populations and settings where they are used. Some also offer age-based norms, providing an empirical method for comparing behavior and emotions against milestones of normative development (Achenbach, 2001). If diagnostic efficiency statistics, such as sensitivity and specificity or diagnostic likelihood ratios are available, then it is possible to combine information from checklist scores with estimates of baseline probability, and other risk factors to come up with a revised probability of diagnosis (Straus, Glasziou, Richardson, & Haynes, 2011). The assessment methods advocated by Evidence Based Medicine (EBM) use Bayesian techniques, packaged in a way that is accessible to clinicians, to integrate the information from test results with other available clinical data.

The application of these methods to the specific problem of diagnosing pediatric bipolar disorder has already shown large effect sizes for changing clinical practice by making estimates more accurate, eliminating a bias toward overestimating the probability of a bipolar diagnosis, and improving the consistency of agreement (i.e., reducing

the range of opinion between clinicians; Jenkins et al., 2011). The cumulative effect is enhanced agreement about the next clinical action to recommend for a given case (Jenkins, Youngstrom, Youngstrom, Feeny, & Findling, 2012).

Various instruments are now available that assess clusters of symptoms related to bipolar disorder. Some of these, such as the Achenbach System of Empirically Based Assessment (Achenbach & Rescorla, 2001), do not include a mania scale, but contain subscales measuring other symptom principal components, such as attention problems, aggressive behavior, and anxious/depressed symptoms, that bipolar disorder influences (Mick, Biederman, Pandina, & Faraone, 2003). Several other checklists were originally written for adults and then tested for use with adolescents (Danielson, Youngstrom, Findling, & Calabrese, 2003; Wagner et al., 2006), or adapted for parents to report about their child's mood and behavior (Gracious, Youngstrom, Findling, & Calabrese, 2002; Wagner et al., 2006; Youngstrom, Findling, Danielson, & Calabrese, 2001). A few were originally conceived and designed for use with pediatric samples (Papolos, Hennen, Cockerham, Thode, & Youngstrom, 2006; Pavuluri, Henry, Devineni, Carbray, & Birmaher, 2006). Though the development of new measures is progress for the field, it also complicates the instrument selection process. With few head-to-head comparison studies, it is difficult to compare the performance of these instruments. It is timely to do a meta-analysis to compare measure performance, and to identify conceptually and clinically meaningful moderators of measure performance.

Potential Moderators of Diagnostic Accuracy of Measures Used With Youths

Several design issues likely complicate interpretation of the literature on pediatric bipolar disorder (PBD) assessment.

Differences in Informant

It is axiomatic that assessment of youths should involve multiple informants (Sattler, 2002): parents and teachers observe the youth in different, important developmental contexts, and they have different implicit expectations for typical youth behavior. Youths have their own perspective on their lives, and have privileged access to their internal states; but they also show large developmental changes in verbal ability, metacognition, and their degree of psychological mindedness, all of which change the reliability and validity of their responses to rating scales (De Los Reyes & Kazdin, 2005). For all of these reasons, the correlation between caregiver, teacher, and youth ratings tends to be only moderate (e.g., $r \sim .2$ to .3) across a broad range of psychopathology constructs (Achenbach, McConaughy, & Howell, 1987; De Los Reyes & Kazdin, 2005). The moderate degree of agreement has a big impact on the definition of clinical "caseness"—a clinical elevation according to one informant will typically be linked with only modest elevations according to other observers (Youngstrom, Findling, & Calabrese, 2003; Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006a). A practical consideration for both researchers and clinicians is deciding how to proceed when informants do not agree. Requiring unanimity among caregivers, teachers, and youths (e.g., using the AND rule) identifies the most impaired cases (Carlson & Youngstrom, 2003) and sharply reduces false positive rates; but it also identifies only a quarter as many cases as would meet the definition of caseness according to any one of the three informants (e.g., using the Boolean OR rule; Youngstrom et al., 2003). This is directly pertinent to the debates during the *Diagnostic and Statistical Manual for Mental Disorder-Fifth Edition* (*DSM–5*) revision process about whether to require impairment in multiple settings as part of the criteria for establishing a manic episode (Leibenluft, 2011; Youngstrom, 2009).

Informant issues are especially salient in the area of pediatric bipolar disorder. Clinicians often give more credence to youth report of internalizing problems, because youths have more direct access to their own subject mood states (Loeber, Green, & Lahey, 1990; Youngstrom et al., 2011). However, whenever studies have compared youth and caregiver report in the same sample, caregiver report has produced larger effect sizes for discriminating cases with bipolar disorder from other conditions (Hazell, Lewin, & Carr, 1999; Wagner et al., 2006; Youngstrom, Findling, & Calabrese, 2004; Youngstrom et al., 2005). This might be because of bipolar disorder also creating substantial externalizing problems, which collateral informants often notice earlier and find more bothersome than the youth. Mania and hypomania include other symptoms—such as pressured speech or flight of ideas—that others find worrisome sooner than the person experiencing them. Parents notice irritable mood at significantly lower levels of mania than the youth, who may notice symptoms of increased energy, hypersexuality, and decreased need for sleep sooner instead (Freeman, Youngstrom, Freeman, Youngstrom, & Findling, 2011). There is evidence with both youths (Youngstrom, Findling, & Calabrese, 2004) and adults (Dell'Osso et al., 2002) that hypomania and mania compromise peoples' insight into their behavior and how it is perceived by others, possibly further undermining the credibility (Youngstrom et al., 2011) and validity of youth report.

Beyond caregiver and youth report, some experts argue that teacher report is important for "corroborating" mania—that manic symptoms are more credible and likely to be more impairing when observed by multiple informants across multiple settings (Carlson, 2011). Conversely, others assert that teacher report should not be included in the decision-making algorithm for diagnosing pediatric bipolar disorder, as it has less validity than caregiver and possibly than youth report and has failed to demonstrate incremental validity for the purpose of predicting diagnosis (Youngstrom, Jenkins, Jensen-Doss, & Youngstrom, 2012). Unfortunately, less work has evaluated teacher report (relative to caregiver and youth report) with regard to pediatric bipolar disorder. Several studies found that the Achenbach Teacher Report Form (TRF) is significantly elevated on multiple scales in the presence of pediatric bipolar disorder compared with ADHD or to healthy controls (Geller, Warner, Williams, & Zimerman, 1998; Hazell et al., 1999). However, the effect sizes tend to be smaller for teacher report than caregiver report, and the effect sizes shrink further when the comparison group is also treatment seeking instead of healthy controls (Youngstrom, Findling, Calabrese, et al., 2004). Although they have not yet been compared head-to-head in the same sample, teacher report on manic symptom scales also produced smaller effect sizes than caregiver report on the same instruments (Youngstrom, Joseph, & Greene, 2008).

Scale Content

Another potential moderator is the item content of the scale. Widely used measures such as the Achenbach System of Empirically Based Assessment (Achenbach & Rescorla, 2001) do not include a mania scale, and they often do not have items assessing symptoms that might be specific to mania, such as elated mood or grandiosity. The omissions reflect the time period when the item pool was written, predating consideration that bipolar disorder might manifest in childhood (Achenbach & Edelbrock, 1983). Subsequent research using these scales found that youths with bipolar disorder showed elevations on multiple clinical syndrome scales (Mick et al., 2003). These scales are often elevated in the context of other diagnoses besides bipolar disorder, indicating that they are not specific to bipolar. Although

there was initial enthusiasm for a "bipolar profile" consisting of elevations on multiple scales, subsequent research found that many cases showing the profile did not meet criteria for bipolar disorder (Diler et al., 2009; Meyer et al., 2009). Other analyses found that the Externalizing score captured most of the diagnostic information from the Achenbach System of Empirically Based Assessment (ASEBA) with regard to potential bipolar disorder, and there was no incremental value in adding the syndrome scores after looking first at Externalizing (Diler et al., 2009; Kahana, Youngstrom, Findling, & Calabrese, 2003; Youngstrom, Findling, Calabrese, et al., 2004).

In contrast, other scales focus on symptoms of mania, either using the *DSM* symptoms as the basis of the items (e.g., Hirschfeld et al., 2000; Pavuluri et al., 2006), or even expanding the item pool to include other clinical features that might be associated with hypomania or mania in addition to the canonized *DSM* symptoms (e.g., Depue et al., 1981; Papolos et al., 2006). These scales are likely to be more diagnostically sensitive to bipolar disorder because they ask directly about the relevant symptoms. They also may be more diagnostically specific to bipolar disorder inasmuch as they also include distinctive symptoms.

Differences in Interview Strategy

In addition to the question of who completes the rating scale to describe the youth's emotions and behavior, it also is vital to consider how we arrive at our diagnoses. Mental health lacks the equivalent of an autopsy or pathology report that can conclusively establish a diagnosis. In a field where a "gold standard" diagnosis is impossible, perhaps the best we can do is a "LEAD" standard—the Longitudinal, Expert evaluation of All Data—including history of development, prior treatment, and response, family history of pathology, and integration of collateral informant perspectives as well as direct observation of behavior (Spitzer, 1983). Many research studies approximately approach the LEAD standard by combining a semistructured interview with expert clinician review and sometimes unstructured interviewing to fill in gaps or probe alternate hypotheses. Semistructured interviews are much less likely to be used in clinical practice because of length, as well as practitioners valuing autonomy (Garb, 1998). As we include fewer additional sources of information in the diagnostic process, we must place greater weight on the remaining ones.

The least common denominator in clinical diagnoses of children is an interview with the primary caregiver. The caregiver is most likely to initiate the referral for outpatient services, and young children are unlikely to have the patience, focus, or metacognition needed to complete many semistructured interviews. If the interview is redesigned to be developmentally appropriate for young children, it is difficult to connect with adolescent and adult interviews or diagnostic nosologies (e.g., Ablow et al., 1999; Wakschlag et al., 2012). However, if the diagnostic formulation is based solely on the caregiver interview, then there is no source of potentially disconfirming information. Several factors can undermine the validity of caregiver report, including the caregiver's own stress or psychopathology (De Los Reyes & Kazdin, 2005; Richters, 1992), seeking disability or educational accommodations ("secondary gains"), or complex interactions around issues with the juvenile justice system or child custody (Sattler, 2002). Even if the youth does not complete a semistructured interview, direct interaction and observation provide key data about mental status, the presence or absence of stereotypic behavior, and a variety of other factors that can change diagnoses (Carlson & Youngstrom, 2011; Morrison, 2007).

The issue of interview informants has prompted much discussion within the field of pediatric bipolar disorder research. Although most research groups gravitated toward using some version of the Kiddie Schedule for Affective Disorders and Schizophrenia (Geller et al., 2001; Kaufman et al., 1997; Orvaschel, 1995) as the core semistructured diagnostic interview (Nottelmann, 2001), some groups relied primarily or solely on parent interviews when the case was a youth younger than 12 years (e.g., Biederman et al., 1995; Dienes, Chang, Blasey, Adleman, & Steiner, 2002; Meyer et al., 2009; Papachristou et al., 2013). Others insisted on also interviewing the youth (e.g., Carlson, Loney, Salisbury, & Volpe, 1998; Findling, Youngstrom, et al., 2005; Geller et al., 1998). When groups reported different rates of comorbid pervasive developmental disorders, anxiety disorders, or family histories of antisocial personality and other parental diagnoses (see Biederman et al., 2003; Geller & Luby, 1997; Kowatch, Youngstrom, Danielyan, & Findling, 2005; Youngstrom et al., 2008, for reviews), it became important to isolate the source of the differences. In addition to differences in the content and organization of mood items in the different interviews used (Galanter, Hundt, Goyal, Le, & Fisher, 2012), differences in training, or differences in ascertainment and referral patterns, interviewing only the parent may inflate the association between caregiver-reported checklists and diagnoses—even when the diagnosis is blind to the checklist and based on semistructured interview (Carlson & Klein, 2014). Unlike factors reviewed above, interviewing only the parent likely affects both sensitivity and specificity by exaggerating the degree of separation between the distributions for those with versus without the diagnosis. Leaning heavily on the caregiver for both the criterion and the predictor will exaggerate the apparent effect size.

Study Design Features Especially Potent in Diagnostic Efficiency Studies

Experts have developed standardized guidelines for reporting and critically evaluating the design features of studies evaluating diagnostic tests (e.g., STARD; Bossuyt et al., 2003a) as well as general reports of empirical studies (American Psychological Association, 2008). Here we will focus on factors that (a) affect the severity of the target condition, thus altering the diagnostic sensitivity; and (b) affect the composition of the comparison group, thereby changing the diagnostic specificity.

Design factors changing the diagnostic sensitivity of a measure. The more severe the illness, the easier it is to distinguish from other conditions. Factors affecting the severity of the target condition include the stage of illness, the severity of the presentation, and the use of broad or narrow target definitions (Zhou, Obuchowski, & McClish, 2002). For bipolar disorder, the variability in mood states further complicates the picture because the same illness may manifest with periods of euthymia, hypomania, mania, dysthymia, depression, or mixed mood presentations. Additionally, unlike most diagnoses, bipolar diagnoses persist even after the person recovers from an episode, technically being coded as "in remission." Therefore, bipolar disorder is heterogeneous—spanning from high functioning people in remission all the way to severely disorganized behavior requiring psychiatric hospitalization. In practice, the severity of illness correlates with participants' recruitment setting: inpatient samples have the highest average degree of mania, community samples the lowest average (Lewinsohn, Klein, & Seeley, 2000; Merikangas et al., 2011), and outpatient samples usually fall in between. All else being equal, mania will be easier than hypomania to tell apart from ADHD or depression. Diagnostic sensitivity of tests will vary as a direct function of the severity of the illness (Zhou et al., 2002).

Similarly, the "broad" versus "narrow" definition of diagnosis plays a prominent role in pediatric bipolar disorder. The narrowest research operational definitions require the presence of elated mood and/or grandiosity, whereas irritable mood would be sufficient using

DSM–IV and *DSM–5* criteria (sometimes characterized as the "intermediate" phenotype; Leibenluft, Charney, Towbin, Bhangoo, & Pine, 2003). At the other extreme, some groups may have relaxed the requirement of distinct episodes of change in mood or energy, potentially stretching the "broad phenotype" to include cases that do not share core features of bipolar illness (cf. Leibenluft, 2011; Papolos, 2003; Wozniak et al., 1995). Consequently, youth diagnosed using broad definitions of PBD are likely to be more difficult to distinguish from other cases than youth diagnosed using narrow criteria.

Another wrinkle within bipolar disorder comes from the diagnoses of cyclothymic disorder and bipolar Not Otherwise Specified (NOS—the term used in *DSM–IV*) or Other Specified Bipolar and Related Disorders (OS-BRD—the *DSM–5* parlance; American Psychiatric Association, 2013). Cyclothymic disorder is rarely used in clinical practice in the United States (Youngstrom, Youngstrom, & Starr, 2005) yet is more common than bipolar I in epidemiological samples (Merikangas & Pato, 2009; Van Meter, Moreira, & Youngstrom, 2011) and is associated with a high degree of impairment in youths (Van Meter, Youngstrom, Demeter, & Findling, 2013; Van Meter, Youngstrom, Youngstrom, Feeny, & Findling, 2011). Similarly, bipolar NOS appears more common than bipolar I in outpatient youth samples, and is associated with a high degree of impairment (Axelson et al., 2006; Findling, Youngstrom, et al., 2005). However, both cyclothymic disorder and bipolar NOS, by definition, have less severe manic symptoms than bipolar I or II and may be harder to identify using manic symptom checklists. Further complicating the matter, both cyclothymia and bipolar NOS progress to bipolar I or II at high rates during prospective follow-up, raising questions about whether these are prodromes or early stages of illness rather than distinct disorders (Hauser & Correll, 2013; Van Meter, Youngstrom, & Findling, 2012; Vieta, Reinares, & Rosa, 2011). This creates tension between internal versus external validity: samples focusing on acutely manic bipolar I presentations will produce higher sensitivity estimates, but the results will generalize less well to applications where the goal is to identify bipolar spectrum disorder or earlier, milder stages of bipolar disorder.

All of these considerations change the diagnostic sensitivity of the test because they change the distribution of scores among the target group. Sensitivity is defined as the percentage of cases having the target diagnosis that also score above a designated threshold on the test of interest. The average score on the scale will be higher if the severity of presentation is more extreme. As the distribution shifts toward higher scores, a larger percentage of people will score above any given threshold, increasing the sensitivity of the test. For our purposes, studies with greater rates of bipolar I, more cases with current manic episodes, or drawing larger percentages from inpatient settings are all likely to have higher average scores on scales intended to detect mania. More subtly, samples including broader definitions of bipolar disorder, or enrolling people in varying states of illness, will tend to have more variation in scores. In addition to altering the sensitivity of the scale, the greater variance within the bipolar group also increases the overlap in score distribution with the comparison group, reducing the scale's diagnostic accuracy.

Design factors changing the diagnostic specificity of a measure. The composition of the comparison group directly affects the diagnostic specificity of the measure. Anything that lowers the mean, or decreases the variability in the distribution of scale scores in the comparison group will increase the effect size and decrease the amount of overlap between the bipolar and nonbipolar score distributions. One common design element that would have this effect is the inclusion of healthy controls. Healthy controls, by definition, will have low scores on any symptom measure. Adding them to the sample will shift the mean lower. More subtly, because healthy controls tend to show a floor effect on clinical measures, they bunch together at the lower end of the scale and increase skew.

Another design element that can affect specificity is excluding cases with diagnoses that mimic aspects of bipolar disorder. Unipolar depression and bipolar depression look quite similar, for example. ADHD has multiple symptoms and features that overlap with symptoms of hypomania and mania, including high activity, distractibility, and impulsivity (Biederman et al., 1998). Oppositional defiant disorder and conduct disorder also entail high degrees of irritable mood, aggressive behavior, and rule-breaking that can look like the mood or impulsive risky behavior of mania (Bowring & Kovacs, 1992). Posttraumatic stress disorder and schizophrenia can produce symptoms that overlap with mania, shading into the more psychotic presentations. Symptom overlap raises the average score on scales where the content includes symptoms that multiple disorders "share." Endorsing the symptoms because of other disorders raises the average score in the comparison group, increasing the percentage of "false positive" results and directly reducing the specificity of tests (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006b; Zhou et al., 2002). Bias may be stronger for scales that mostly contain nonspecific items, such as irritability and distraction.

If the research design exaggerates diagnostic specificity, a weak test could appear better than a stronger one evaluated in a more generalizable sample. Then subsequent applications of the test, under more clinically realistic conditions would produce systematically higher rates of false positives (the converse of lower diagnostic specificity), creating upwardly biased posterior probabilities. Flawed research designs could reintroduce the same bias that rating scales were intended to fix.

Nested Effect Sizes and Technical Issues in Establishing Relative Superiority

Another issue is trying to determine whether some measures perform significantly better than others, after accounting for differences because of informant or research design. It is possible to test the difference between effect sizes from different samples, comparing the discrepancy to what would be expected under the null hypothesis of no difference beyond sampling error (Viechtbauer, 2007). More statistically powerful tests are possible when the effect sizes come from the same sample (Venkatraman, 2000), and a few reports have already directly tested performance differences between measures in the same sample (Youngstrom, Findling, Calabrese, et al., 2004; Youngstrom et al., 2006a). These analyses control for illness severity, comparison group composition, interviewer training, and a host of other factors that could differ between studies (Zhou et al., 2002). Unfortunately, there is no "master linking sample" that compares all of the contending measures against each other head to head.

However, a mixed effects meta-analysis model can account for the fact that some measures may be confounded with design features in the available literature (e.g., if only one research group has published results with a particular measure, then it will be harder to tease apart characteristics of the measure from the set of design factors used by the group). Multivariate meta-analyses can disentangle the effects of design artifact from the differences between measures, allowing direct comparison of measure performance (Viechtbauer, 2010). Clinicians would want to know which measure to use for high stakes decisions, and researchers could enhance studies by switching to the more valid measure.

Research Questions/Hypotheses

We expected the average effect size to be large, reflecting a big standardized difference in mean scores for those with bipolar versus other conditions. However, we also expected the effect sizes to show significant heterogeneity, and we had specific hypotheses about moderating variables. Based on the few prior within-sample comparisons, we predicted that caregiver report would show larger average effect sizes than youth or teacher report. We also hypothesized that scale content would matter: measures that ask about symptoms more specific to mania should show larger overall effect sizes than measures that focus on externalizing symptoms in general, or that combine components originally designed to assess depression, attention problems, and aggressive behavior (Althoff, Ayer, Rettew, & Hudziak, 2010; Mick et al., 2003). A third hypothesis was that studies that only directly interviewed the caregiver would produce larger effect size estimates than those that included direct interview and observation of the youth as part of the criterion diagnosis, because of shared source variance. A fourth hypothesis was that the use of more "distilled" samples that identify more homogeneous and symptomatic cases of bipolar, exclude diagnoses that frequently are difficult to distinguish from bipolar, or that include healthy controls in the comparison group, would yield much larger effect sizes.

We hypothesized that all moderators would remain significant when entered together in the regression models. If "distilled design" was a significant moderator, we would give primacy to the "nondistilled" estimates of effect size as more clinically generalizable, and treat them as the main focus of discussion. Similarly, if interview strategy (including the youth vs. relying only on the caregiver) moderated results, then the results based on integrated interviews would take precedence, as they would be less affected by shared source variance. We explored whether there were significant differences between scales after controlling for moderators, but anticipated that the variability between samples, combined with the number of multiple comparisons, would make those results tentative. Sensitivity analyses examined whether results changed substantively after controlling for quality of design (following the scheme used in Kowatch et al., 2005) or quality of reporting (using the recommended QUADAS-2 tool developed to operationalize the STARD Guidelines; Whiting et al., 2011).

Method

Inclusion and Exclusion Criteria

Studies were included if they reported (a) cases with a diagnosis of a bipolar spectrum disorder made via a structured or semistructured interview; (b) as well as a comparison group; (c) with both groups completing the same checklists assessing manic, hypomanic, or externalizing symptoms; (d) with data reported for participants 18 years or younger. Cases could be drawn from clinical or community samples. Exclusion criteria included having fewer than 10 cases with bipolar diagnoses (per Kraemer, 1992, to provide reasonably stable estimates of diagnostic sensitivity; e.g., excluding Reichart et al., 2005), not including a rating scale (e.g., Henin et al., 2007), not publishing results in an English format (note that we did not find any studies that had usable effect sizes that had been published in other languages), only having data for the bipolar group and no comparison group (e.g., Wilens et al., 2003), only reporting clinical diagnoses based on chart review or unstructured interviews (e.g., Youngstrom et al., 2005). We limited the search period to 1993 and later so that the *DSM–IV* criteria would be available and used. Functionally, there were no group comparison studies published before then on the topic anyway; only case reports (e.g., Anthony & Scott, 1960). There were no geographical or cultural restrictions. Studies with adult samples were included only if they reported sufficient information about the subset of cases 18 years and younger (cf. de Sousa Gurgel, Reboucas, Negreiros de Matos, Carneiro, & Gomes de Matos e Souza, 2012; Meyer et al., 2007; Miller, Johnson, Kwapil, & Carver, 2011; Zaratiegui et al., 2011—all of which were reviewed in Waugh et al., 2014, but failed the inclusion criteria here). We excluded effect sizes and studies where the groups were not defined by diagnostic interviews, but instead by proxy definitions of bipolarity based on rating scales, such as the Child Behavior Checklist (CBCL) proxy (e.g., Doerfler, Connor, & Toscano, 2011; Mbekou, Gignac, MacNeil, Mackay, & Renaud, 2014), elevated scores on a parent-reported mania scale (Carlson & Kelly, 1998), or "corroborated" mania reported by multiple informants on the same rating scale (e.g., Carlson & Youngstrom, 2003). These scenarios involve criterion contamination, where the scores on the measure contributed directly to the determination of the criterion "diagnosis" definition (Bossuyt et al., 2003b; Zhou et al., 2002). Per *DSM–IV*, bipolar spectrum diagnoses could include bipolar I, bipolar II, cyclothymic disorder, and bipolar Not Otherwise Specified (NOS). All studies included in the analysis reported that they used *DSM–IV* criteria, but the publications did not report effect sizes separately for the different bipolar diagnoses, so it was not possible to estimate effect sizes for each type of bipolar disorder.

Moderator Definitions

We created a coding manual in Microsoft Excel, where the variable names, definitions, value labels, and examples were in rows or comment boxes next to the coding area. In addition to publication year, country of data collection, clinical setting (epidemiological/general community, outpatient, acute tertiary setting), and variables necessary for coding study design and reporting quality (detailed below), we also coded several potential moderator variables.

Informant. For each effect size, we coded whether the informant completing the checklist or scale was the caregiver (including foster parents or custodial relatives, although in the vast majority of cases across all samples it was the biological mother), the teacher, or the youth. Analyses used dummy codes with caregiver as the reference category.

Type of scale. For each effect size, we coded whether the scale contained symptoms specific to mania versus comprising items or subscales originally designed to measure other pathology. For example, the "bipolar profile" from the ASEBA instruments (Achenbach & Rescorla, 2001) consists of a combination of the Aggressive Behavior, Attention Problems, and Anxious/Depressed scales. There is no "mania" scale on the ASEBA; the manic items it contains are those that overlap with other disorders, and thus factor analyses assigned them to other subscales. The meta-analyses used a dummy code that defined nonspecific scales as the reference category (testing whether there was an advantage in using scales with more mania-specific content).

Interview strategy. For each sample, we coded whether the criterion diagnoses derived from interviews solely with the primary caregiver versus also involving direct interview of the proband youth. One study also included interview with the teacher on an inpatient/residential unit as an additional source (Carlson et al., 1998). We included this study in the "not relying solely on the caregiver" category.

Distilled sample design. This dichotomous variable coded whether the original study used a design likely to inflate the observed effect sizes. This was coded "yes" if the sample included healthy controls as part of the comparison group, lowering the mean score for the comparison group and also potentially lowering the *SD*. It also was

coded yes if the design excluded diagnoses likely to share symptoms similar to those characteristic of bipolar disorder, such as unipolar depression, ADHD, conduct disorder, or psychosis. Many studies of phenomenology of bipolar disorder relied on healthy controls or groups with ADHD but excluded comorbid mood disorder as comparison conditions.

Search Strategies

As recommended in PRISMA (Liberati et al., 2009), we consulted with a social sciences reference librarian while designing and revising the search strategy. Reference and citation databases searched included PubMed, PsycINFO, SSCI, ERIC, and GoogleScholar. We piloted the search protocol, consulted with a reference librarian, and implemented the revised protocol. Either PubMed or PsycINFO indexed all of the published reports that met inclusion criteria. Search terms were: (Pediatric OR juvenile OR child* OR adolescen*) AND ("bipolar disorder" OR mani* OR cyclothymi*) AND [(Sensitivity AND Specificity) OR comparison]. Review articles and chapters were checked for additional sources. This generated 1342 hits in PsycINFO, and 4,094 hits in PubMed when the search was updated on September 1, 2014. We pulled hits into a RefWorks database, where we could sort them and annotate them to track disposition. Four relevance judges completed the search training (including review of guidelines and a session of orientation and consultation with a reference librarian about search optimization) and then conducted and reviewed the searches. A content expert (EAY) reviewed all ambiguous cases and instances of disagreement. After reviewing titles and abstracts, and initial elimination of multiple publications using the same dataset, we retrieved 69 articles for detailed review and coding. We examined the reference lists in all studies that met inclusion criteria, along with scrutinizing the bibliographies of recent reviews (Geller & DelBello, 2003; Johnson, Miller, & Eisner, 2008; Mick et al., 2003; Waugh, Meyer, Youngstrom, & Scott, 2014; Youngstrom, 2007). The Mick et al. (2003) article identified two additional samples meeting inclusion criteria (Biederman et al., 1996), and a chapter in one edited volume provided sufficient information to add another sample and effect size (Lewinsohn, Seeley, & Klein, 2003). A review of articles found five datasets where the article captured by the search did not include sufficient information, but a second article by the same group included the necessary information (Doerfler, Connor, & Toscano, 2011; Henry, Pavuluri, Youngstrom, & Birmaher, 2008; Lee et al., 2014). We did not locate any primary reports published in languages other than English, although some reports published in English language journals gathered data using translated versions of measures into Korean/Hangul (Lee et al., 2014), French (Miguez et al., 2012), and Dutch (Papachristou et al., 2013). The final dataset included 25 distinct reports reporting 27 samples (two reports published data on two samples). The initial PubMed search identified 12 of 25 usable sources (48% search sensitivity) and indirectly identified 3 more samples (60% search sensitivity, broadly defined); PsycINFO identified 11 sources directly (44% search sensitivity) and 5 more indirectly (64% search sensitivity, broadly defined). The low search sensitivity is partly an artifact of our decision to include studies that reported sufficient statistics even if they did not report diagnostic sensitivity and specificity in the article, as few research groups have used receiver operating characteristic analyses in this literature until recently. In three cases our group obtained access to the primary data and estimated effect sizes directly from the raw data. Figure 1 shows the flow diagram for the search process.

Coding Procedures

Coders were undergraduate psychology majors, doctoral students, and the senior investigator. Training included reading methodology articles (QUADAS, PRISMA, STARD; Bossuyt et al., 2003a; Liberati et al., 2009; Whiting et al., 2011), sample meta-analyses focused on pediatric bipolar disorder (Kowatch et al., 2005; Van Meter et al.,

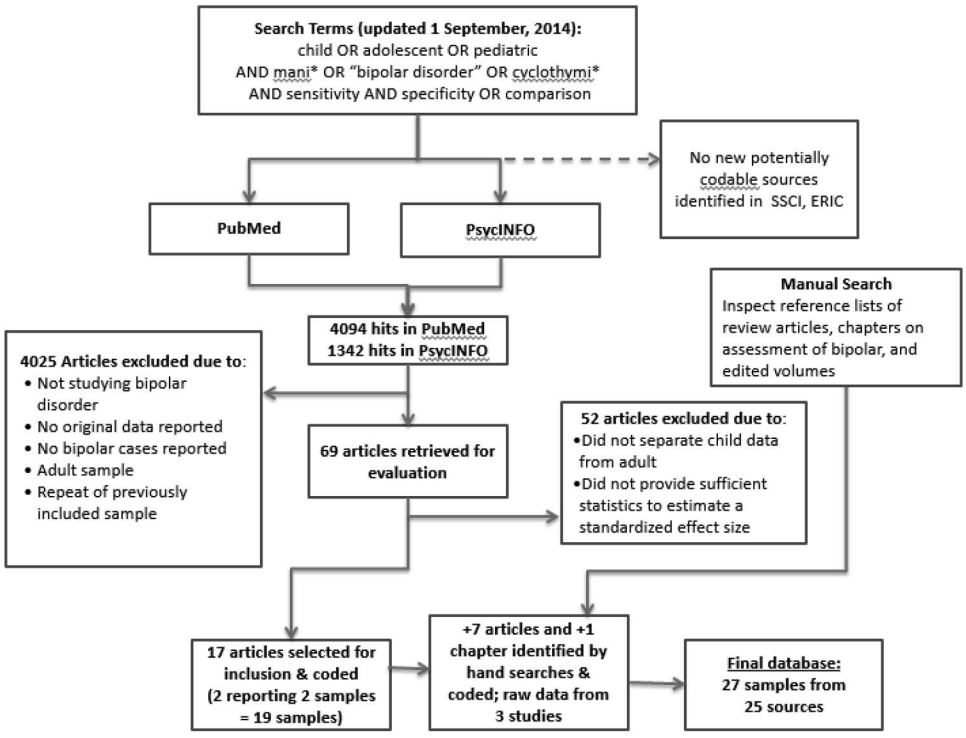


Figure 1. Flow diagram of search strategy and final sample.

2011), orientation to diagnostic efficiency statistics (Youngstrom, 2014), and then coding two articles and comparing scores to those of a content expert, resolving discrepancies and clarifying concepts. We double-coded all studies for effect sizes, moderator variables, and reporting quality. Some articles reported sufficient statistics to estimate the effect size several different ways, contributing to small discrepancies in effect sizes estimates if the coders used different methods. The content expert reviewed all discrepancies and assigned a final code after perusing the source material, using the method that made fewest distributional assumptions to estimate the effect size. In three cases, the raw data were available and analyzed to provide more extensive information than provided in the primary publication, which often was a preliminary report.

Quality ratings. We used two systems to code the quality of the study design and reporting. The first was based on a prior meta-analysis of pediatric bipolar disorder (Kowatch et al., 2005). It assigned points for adequate sample size ($N > 30$), interviewing both caregiver and youth (vs. one informant only), using a formal consensus process, following *DSM* criteria, including spectrum diagnoses (e.g., cyclothymic disorder, and bipolar NOS), recording comorbid diagnoses, and systematically asking about lifetime episodes. Higher scores indicated more comprehensive assessment of the bipolar phenotype.

The second quality coding scheme was the Quality Assessment of Diagnostic Accuracy Studies, Version 2 (QUADAS-2; Whiting et al., 2011), a standardized coding protocol for articles reporting diagnostic efficiency results that operationally defines coding criteria for the STARD guidelines. We retained all of the QUADAS-2 items, though some were rarely reported in the target articles. Both the Kowatch score and the QUADAS-2 score served as covariates in sensitivity analyses to determine whether design or reporting quality accounted for significant variance in effect sizes, and to determine whether the moderators of interest survived correction for these design quality factors.

Rater reliability. Interrater agreement was good (ICC for absolute agreement $> .87$ for demographics and moderator variables, $> .95$ for effect size metrics, and $> .80$ for quality ratings). The most likely source of disagreement was when raters selected different formulae for estimating effect sizes, or when one coder was aware of algebraic methods that could transform reported information into something that could be extracted and coded, and the other rater had coded the parameter as missing.

Statistical Methods

We used Hedges' g, a standardized mean difference that corrects Cohen's d for a slight upward bias in small samples, as our summary effect size (Lipsey & Wilson, 2001). There are three advantages to using standardized mean difference for the purposes of meta-analysis: (a) the studies reviewed more often reported Cohen's d than area under the curve (AUC); (b) meta-analytic techniques are more highly developed for standardized mean difference than combining AUCs; and (c) analysis of sensitivity and specificity create technical challenges avoided by focusing on other metrics (Hasselblad & Hedges, 1995; Zhou et al., 2002). We used standard formula to convert sufficient statistics into g (see Lipsey & Wilson, 2001, for list and formulae). AUC converts directly to Cohen's d, and then to g. If only sensitivity and specificity were reported, these could be converted to an AUC estimate (Hasselblad & Hedges, 1995), and then to a d and finally a g. Sample means, SDs, and n for the bipolar and comparison group also were sufficient for direct estimation of g. Study variance estimate calculations followed standard methods (Viechtbauer, 2010). All estimates used inverse variance weighting, and we report 95% confidence intervals (CIs) for the weighted effect sizes.

Most studies reported multiple relevant effect sizes. The nesting of several effect sizes in the same sample could occur because of the use of multiple informants (e.g., caregiver, youth, or teacher report), comparison of multiple scales in the same study, or reanalysis of data to examine the influence of sampling design (distilled or not) on effect size estimates. When studies reported multiple scales from the same measure, analyses used a single estimate: Externalizing was the preferred CBCL scale because it has tied or outperformed "bipolar profiles" in multiple samples (Diler et al., 2009; Kahana et al., 2003). We used brief versions instead of full-length versions when both were reported because they are more likely to be used in practice. The *metafor* package (Viechtbauer, 2010b) in R (R Core Team, 2014) was the platform for all analyses, as it is one of the few meta-analysis programs that currently handles nested effect sizes within the same sample (Viechtbauer, 2010a), allowing us to test key moderator variables.

Analyses used mixed metaregression models. We had several hypothesis-driven moderators of interest, but also want to preserve generalizability, so a mixed approach was best (Viechtbauer, 2010a). We examined each moderator separately, but also created a fully augmented model to test whether each moderator showed a unique incremental effect. Cochran's Q tested homogeneity of effect sizes, along with graphical methods (e.g., forest plots—see Figure 2). Non-**F2** significant Q values indicate little heterogeneity beyond sampling error. We used a mixed model extension of Egger's test for publication bias, although we expected publication bias to be low because diagnostic efficiency requires large effect sizes, making statistical significance a relatively low bar to exceed. We examined standardized residuals from the fitted models, instead of funnel plots, as a way of testing for influential outliers while accounting for the nested structure of the data (Viechtbauer, 2010a).

Similarly, the Meta-Analysis Reporting Standards (MARS; American Psychological Association, 2008) suggest estimating power when conducting meta-analyses. Power exceeded 99.9% to reject the null hypothesis of $g \sim 0$, because effect sizes need to be $g > .5$ to begin to provide diagnostically useful information, and preferably much larger (Hummel, 1999). To account for nesting, we bracketed power estimates by using the number of independent samples (27) as a low end and the number of effect sizes as the high end. Power to detect moderate heterogeneity (e.g., values of .67) was between .64 and .91 based on 27 independent samples and 63 disaggregated effects, respectively (Hedges & Pigott, 2001). Power was between .86 and .99 for large heterogeneity. We used outlier diagnostics to identify influential cases (Viechtbauer, 2010a), and we conducted robustness sensitivity analyses to examine their effects on parameter estimates.

Results

Figure 1 presents a flow diagram showing the search process. We identified 27 distinct samples from 25 reports published between 1995 and 2014, contributing 63 effect sizes. Of the effect sizes, 38 used caregiver report on a total of 10,232 youths between the ages of 5 and 18 years: 1,719 with research interview diagnoses of bipolar disorder, 3,150 healthy controls or youths from the general community, and 5,363 with other disorders besides bipolar spectrum diagnoses. Youth report generated 14 effect sizes (based on 448 youths with bipolar diagnoses, 1,028 healthy youths, and 1,542 with other diagnoses), and teacher report had 11 effect sizes (based on 377 cases with bipolar diagnoses, 58 healthy youths, and 855 with other diagnoses). All child and teacher effect sizes were nested within subsets of caregiver data with the exception of the Lewinsohn et al. (2003) chapter, which only

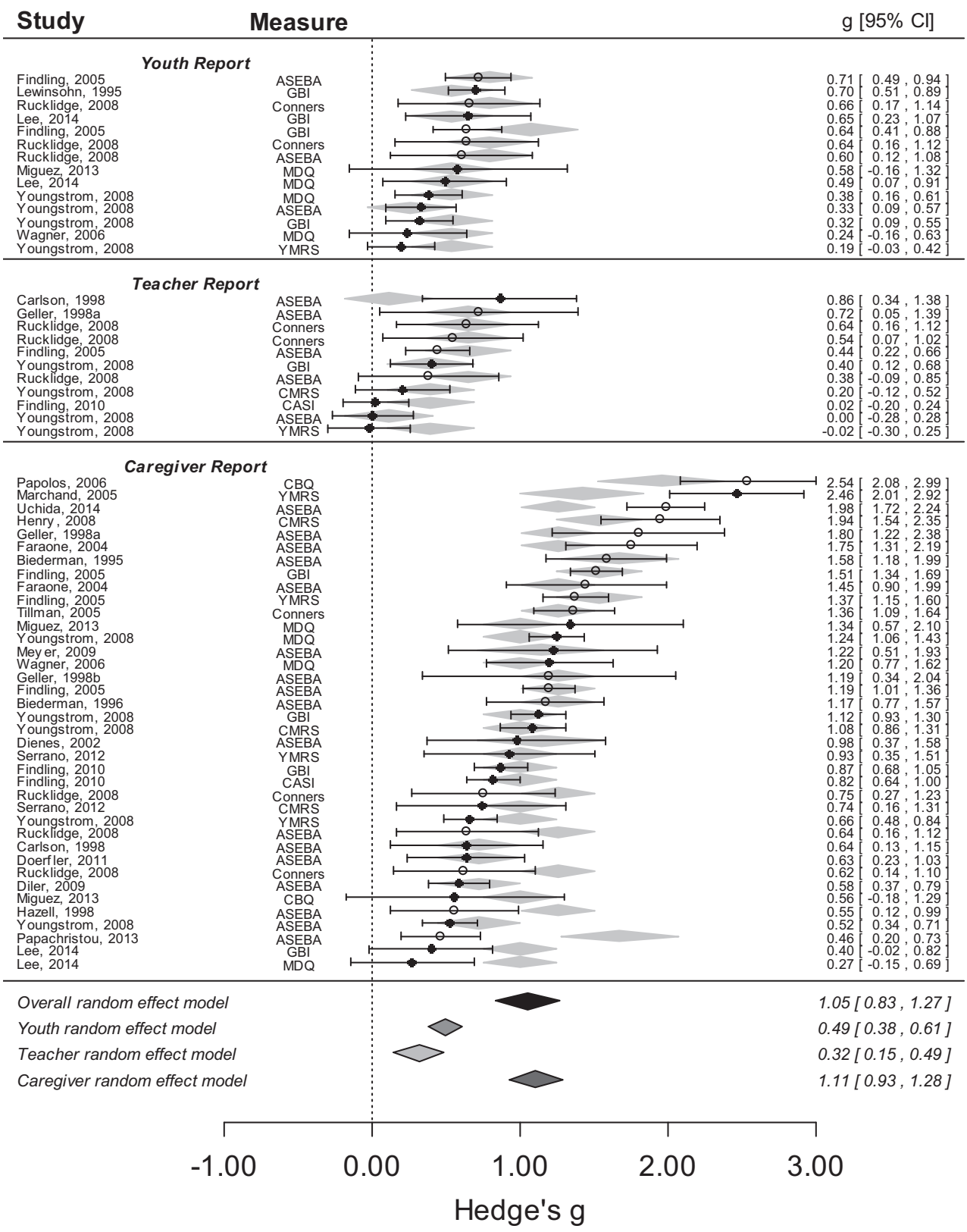


Figure 2. Forest plot of effect sizes, sorted by informant and then descending order of effect size. Filled dots indicate a generalizable, nondistilled design; open dots mark a distilled design. Note that all standard errors are based on a random effects, disaggregated model; whereas primary analyses use multivariate random effects model. Gray polygons indicate the predicted effect size. The Achenbach System of Empirically Based Assessment (ASEBA) and Conners do not have a mania scale, although they have items that could be nonspecific manic symptoms embedded on other scales. GBI = General Behavior Inventory; MDQ = Mood Disorders Questionnaire; YMRS = Young Mania Rating Scale; CMRS = Child Mania Rating Scale; CASI = Child and Adolescent Symptom Inventory; CBQ = Child Bipolar Questionnaire.

included youth report. In terms of other candidate moderator variables, 14 of the 27 samples (52%) used distilled sample designs, and 31 of 63 effect sizes (49%) were based on scales with mania symptom content. Effect sizes came from samples in seven countries, most from the United States, but with each informant contributing effect sizes in at least two countries (see Table 1 for a summary of sample-level characteristics). Eight different checklists contributed effect sizes: the ASEBA contributed 25 effect sizes; the General Behavior Inventory (GBI; Depue et al., 1981) contributed 9; the Mood Disorders Questionnaire (MDQ; Hirschfeld et al., 2000) added 8; the Conners (1999) had 7, the questionnaire version of the Young Mania Rating Scale (YMRS; Gracious et al., 2002) added 6, the Child Mania Rating Scale (CMRS; Pavuluri et al., 2006) provided 4, and the Child and Adolescent Symptom Inventory (CASI; Gadow & Sprafkin, 1994, 1997) and Child Bipolar Questionnaire (CBQ; Papolos et al., 2006) had 2 each. Table 2 reports the effect sizes, along with the moderator variables and other statistics at the level of the effect size. Because some effect sizes used full-length scales and others used short forms, Table 2 also reports the number of items constituting the scale for each effect size. If effect sizes were based on different sample sizes, because of changes in informant or missing data, then we used the *N* for each effect size rather than a single estimate or weight for the whole sample. Figure 2 displays the forest plot for the raw effect sizes, sorted by informant and magnitude of effect, and using shading to show whether the sample used a distilled design.

Assessment of study quality. We used two a priori measures of study quality. The Kowatch system rated design features important for the investigation of pediatric bipolar disorder. All studies reported sufficient information to code all of the Kowatch criteria (except for one item from Hazell et al., 1999). Scaled as percent of maximum possible score, study quality ranged from 50 to 100%, with an average of 83%. The overall quality of the studies included was good in terms of using semistructured interviews, implementing *DSM* criteria, capturing comorbid and confounding diagnoses, and other features that enhance confidence in the robustness of findings.

In terms of the quality of *reporting* results, scores ranged from 45 to 95%, with an average of 73% on the QUADAS-2. Published reports often omitted QUADAS-2 elements: only 12 samples clearly reported the time interval between the diagnostic interview and gathering the rating scales; only 20 clearly specified whether or not the diagnoses were blind to the rating scales; only 21 made clear whether the rating scale was interpreted without prior knowledge of the diagnosis; and only 9 of 25 studies reported all suggested elements. No study included a flow diagram.

Overall Summary of Effect Sizes

We used a multivariate metaregression (rma.mv in *metafor*), modeling the nesting of the effect sizes in the 27 samples and treating both the within study and between study variance estimates as random effects. The overall estimate of effect size was $g = 1.05$. There was tremendous heterogeneity, Cochran's $Q(62\ df) = 738.25$, $p < .00005$. There were substantial variance components both for the within samples nesting of effect sizes (level 1 in a hierarchical linear model

Table 1
Summary of Sample-Level Characteristics of Studies Included in Meta-Analysis

	Study level characteristics					Quality ratings	
Study	Nested effects	Country	Mean age	Setting type	Interviewed	Kowatch total	QUADAS-2 total
Biederman et al. (1995)	1	USA	8.7	Outpatient	Parent	64%	74%
Biederman et al. (1996)	1	USA	11.0	Outpatient + Community	Parent, Child	100%	58%
Carlson, Loney et al. (1998)	2	USA	8.0	Outpatient	Parent, Child, Teacher	86%	68%
Dienes et al. (2002)	1	USA	11.9	At risk	Parent	86%	66%
Diler et al. (2009)	1	USA	9.4	Academic	Parent, Child	86%	68%
Doerfler et al. (2011)	1	USA	10.7	Outpatient	Parent, Child	79%	66%
Faraone et al. (2005)	1	USA	10.5	Outpatient	Parent, Child	79%	84%
Faraone et al. (2005)	1	USA	11.6	Outpatient	Parent, Child	79%	84%
Findling et al. (2005)[a]	6	USA	11.3	Outpatient	Parent, Child	100%	92%
Findling et al. (2010)[a]	3	USA	6.0	Outpatient	Parent, Child	100%	92%
Geller et al. (1998) (males)	2	USA	8.5	Outpatient	Parent, Child	93%	66%
Geller et al. (1998) (females)	1	USA	8.5	Outpatient	Parent, Child	93%	66%
Hazell et al. (1999)	1	Australia	11.0	Community	Parent, Child	50%	47%
Henry et al. (2008)	1	USA	10.3	Community	Parent, Child	71%	87%
Lee et al. (2014)[b]	4	South Korea	13.7	Outpatient + Community	Parent, Child	86%	55%
Marchand et al. (2005)	1	USA	10.7	Community	Parent	79%	55%
Meyer et al. (2009)	1	USA	13.3	At risk	Parent	86%	76%
Miguez et al. (2012)[c]	3	Switzerland	16.0	6 Outpatient/ 2 Inpatient	Parent, Child	79%	95%
Papachristou et al. (2013)[d]	1	Netherlands	16.0	Community	Parent	57%	58%
Papolos et al. (2006)	1	USA	11.0	Outpatient	Parent	93%	74%
Rucklidge (2008)	9	New Zealand	15.4	Community	Parent, Child	86%	45%
Serrano et al. (2011)[e]	2	Spain	11.0	Outpatient	Parent, Child	86%	76%
Tillman & Geller (2005)	1	USA	10.5	Outpatient	Parent, Child	79%	79%
Uchida et al. (2014)	1	USA	10.3	Outpatient + Community	Parent, Child	79%	82%
Wagner et al. (2006)	2	USA	14.5	Outpatient	Parent, Child	79%	76%
Youngstrom et al. (2005)[a]	13	USA	10.9	Outpatient	Parent, Child	100%	95%

[a] Effect sizes estimated using raw data for these samples. [b] Scales translated to Hangul (Korean). [c] Scales translated to French. [d] Scales translated to Dutch. [e] Scales translated to Castilian Spanish. QUADAS-2 = Quality Assessment of Diagnostic Accuracy Studies, Version 2.

Table 2
Effect Size Level Characteristics and Moderators

Study	Nested effects	Participants			Predictor			Moderators			Effect size	
		N bipolar	*N* healthy	*N* other	Measure	Scale	Items	Informant	Mania scale	Distilled design	Hedge's *g*	Weight
Biederman et al. (1995)	1	31	77	120	ASEBA	Externalizing	35	Caregiver		Y	1.58	.04
Biederman et al. (1996)	1	30	107	99	ASEBA	Externalizing	35	Caregiver		Y	1.17	.04
Carlson et al. (1998)	2	23	0	46	ASEBA	Externalizing	35	Caregiver			.64	.07
					ASEBA	Externalizing	45	Teacher			.86	.07
Dienes et al. (2002)	1	16	18	24	ASEBA	Externalizing	35	Caregiver			.98	.09
Diler et al. (2009)	1	157	0	228	ASEBA	Externalizing	35	Caregiver			.58	.01
Doerfler et al. (2011)	1	27	0	249	ASEBA	Externalizing	35	Caregiver			.63	.04
Faraone et al. (2005)	1	22	242	229	ASEBA	Externalizing	41	Caregiver		Y	1.75	.05
Faraone et al. (2005)	1	14	109	287	ASEBA	Externalizing	41	Caregiver		Y	1.45	.08
Findling et al. (2005)[a]	6	291	30	346	GBI	10M	10	Caregiver	Y		1.51	.01
		263	30	309	ASEBA	Externalizing	35	Caregiver			1.19	.01
		174	30	178	YMRS	Total	11	Caregiver	Y		1.37	.01
		151	30	154	ASEBA	Externalizing	32	Teacher			.44	.01
		114	30	174	GBI	Hypomanic/ biphasic	28	Youth	Y		.64	.01
		139	30	182	ASEBA	Externalizing	32	Youth			.71	.01
Findling et al. (2010)[a]	3	162	0	530	CASI	Mania	9	Caregiver	Y		.82	.01
		162	0	530	GBI	10M	10	Caregiver	Y		.87	.01
		100	0	367	CASI	Mania	9	Teacher	Y		.02	.01
Geller et al. (1998)	2	13	0	30	ASEBA	Externalizing	32	Teacher		Y	.72	.12
(males)		27	0	38	ASEBA	Externalizing	35	Caregiver		Y	1.80	.09
Geller et al. (1998)												
(females)	1	12	0	13	ASEBA	Externalizing	35	Caregiver		Y	1.19	.19
Hazell et al. (1999)	1	25	27	99	ASEBA	Externalizing	35	Caregiver		Y	.55	.05
Henry et al. (2008)	1	50	50	50	CMRS	Mania short form	10	Caregiver	Y	Y	1.94	.04
Lee et al. (2014)	4	25	125	73	GBI	10M	10	Caregiver	Y	Y	.40	.05
					MDQ	Raw score	13	Caregiver	Y	Y	.27	.05
					GBI	Hypomanic/ biphasic	28	Youth	Y	Y	.65	.05
					MDQ	Raw score	13	Youth	Y	Y	.49	.05
Lewinsohn et al. (1995)	1	115	845	749	GBI	Short hypomania	12	Youth	Y		.70	.01
Marchand et al. (2005)	1	64	0	66	YMRS	Total	11	Caregiver	Y		2.46	.05
Meyer et al. (2009)	1	9	42	46	ASEBA	Externalizing	41	Caregiver			1.22	.13
Miguez et al. (2012)	3	8	0	68	CBQ	Total	65	Caregiver	Y		.56	.14
					MDQ	Raw score	15	Caregiver	Y		1.34	.15
					MDQ	Raw score	15	Youth	Y		.58	.14
Papachristou et al. (2013)	1	56	1201	973	ASEBA	Externalizing	35	Caregiver		Y	.46	.02
Papolos et al. (2006)	1	76	38	21	CBQ	Total	84	Caregiver	Y	Y	2.54	.05
Rucklidge (2008)	9	25	28	29	ASEBA	Externalizing	35	Caregiver		Y	.64	.06
					Conners	Inattentive	9	Caregiver		Y	.75	.06
					Conners	Hyper/impulsive	9	Caregiver		Y	.62	.06
					ASEBA	Externalizing	32	Teacher		Y	.38	.06
					Conners	Inattentive	9	Teacher		Y	.64	.06
					Conners	Hyper/impulsive	9	Teacher		Y	.54	.06
					ASEBA	Externalizing	32	Youth		Y	.60	.06
					Conners	Inattentive	9	Youth		Y	.66	.06
					Conners	Hyper/impulsive	9	Youth		Y	.64	.06
Serrano et al. (2011)	2	28	0	86	YMRS	Total	11	Caregiver	Y		.93	.09
					CMRS	Total	21	Caregiver	Y		.74	.09
Tillman & Geller (2005)	1	93	94	81	Conners	Short form	2	Caregiver		Y	1.36	.02
Uchida et al. (2014)	1	140	117	83	ASEBA	Externalizing	41	Caregiver		Y	1.98	.02
Wagner et al. (2006)	2	41	0	63	MDQ	Raw total	15	Caregiver	Y		1.20	.05
					MDQ	Raw total	15	Youth	Y		.24	.04
	13	141	0	640	ASEBA	Externalizing	35	Caregiver			.52	.01
		147	0	659	YMRS	Total	11	Caregiver	Y		.66	.01
Youngstrom et al. (2005)[a]		106	0	442	CMRS	Mania short form	10	Caregiver	Y		1.08	.01
		150	0	667	GBI	10M	10	Caregiver	Y		1.12	.01
		152	0	667	MDQ	Raw total	13	Caregiver	Y		1.24	.01
		65	0	220	YMRS	Total	11	Teacher	Y		−.02	.02
		65	0	229	ASEBA	Externalizing	32	Teacher			.00	.02
		49	0	162	CMRS	Total	15	Teacher	Y		.20	.03
		65	0	226	GBI	10M	10	Teacher	Y		.40	.02
		95	0	378	YMRS	Total	11	Youth	Y		.19	.01
		93	0	373	GBI	Hypomanic/ biphasic	28	Youth	Y		.32	.01
		86	0	362	ASEBA	Externalizing	32	Youth			.33	.01
		95	0	376	MDQ	Raw total	13	Youth	Y		.38	.01

[a] Hedges' *g* is an effect size that adjusts Cohen's *d* to correct for upward bias in small samples. ASEBA = Achenbach System of Empirically Based Assessment; GBI = General Behavior Inventory; MDQ = Mood Disorders Questionnaire; YMRS = Young Mania Rating Scale; CMRS = Child Mania Rating Scale; CASI = Child and Adolescent Symptom Inventory; CBQ = Child Bipolar Questionnaire.

Table 3
Tests of Homogeneity and Estimates of Random Effects Variances Between Effect Sizes (Level 1) and Between Samples (Level 2) for Multivariate Meta-Regression Models Using Maximum Likelihood Estimation

Model	Level 1 variance	Level 2 variance	*Q* Residual (*df*)	*Q* Model (*df*)
No moderators	.131	.197	738.25 (62)****	—
Moderator: Informant	.042	.209	401.58 (60)****	53.84 (2)****
Moderator: Mania scale content	.117	.223	728.38 (61)****	1.98 (1)$^{n.s.}$
Moderator: Only parent interviewed	.136	.128	707.65 (61)****	5.80 (1)**
Moderator: Distilled design	.133	.126	625.24 (61)****	6.64 (1)**
Moderators: All simultaneously	.034	.119	279.55 (57)****	84.03 (5)****
All moderators plus Kowatch and QUADAS2 Quality	.032	.106	273.63 (55)****	92.20 (7)****
Compare measures against ASEBA, controlling for informant, distilled design, and only parent interviewed	.022	.143	253.85 (51)****	111.17 (11)****

Note. The augmented model including all moderators, but not the quality ratings, produced the best fit. Although "mania scale content" was not a significant moderator by itself, it became significant after controlling for other moderators. Because quality ratings did not improve model fit, subsequent analyses are based on the model with all other moderators, but not quality. Results controlling for quality did not change substantively. ASEBA = Achenbach System of Empirically Based Assessment. QUADAS-2 = Quality Assessment of Diagnostic Accuracy Studies, Version 2.
** $p < .005$. **** $p < .00005$, two-tailed.

conceptual framework)—sigma^2 = .13, as well as between samples (Level 2)—sigma^2 = .20. This became the baseline model for exploration of moderators and covariates. Table 3 reports the variance estimates and Cochran's *Q* for this and the subsequent augmented multivariate metaregression models.

Moderator Analyses

Informant: Caregiver versus youth or teacher report. Our primary moderator of conceptual interest was the informant who completed the scale. Multivariate metaregression used two dummy codes, comparing youth versus caregiver and teacher versus caregiver. The multivariate framework allowed simultaneous inclusion of all effect sizes and studies in the analysis versus needing to run analyses separately by informant on different subsets (thus number of studies and number of cases is consistent across all moderator analyses).

Informant type explained a significant amount of the heterogeneity, $Q(2\ df) = 53.84$, $p < .00005$. Furthermore, the within-study variance estimate dropped to sigma^2 = .04 when including informant in the model (see Table 3). The parameter estimates indicated that caregiver report produced the largest effect size, $g = 1.20$, with youth report averaging $g = -0.48$ lower, and teacher report $g = -0.65$ lower (all $p < .00005$).

Mania scale content. Scales with mania-specific content should reduce false positive response rates in other diagnostic groups, increasing the effect size. Multivariate metaregression using dummy-coded mania scale content as the sole moderator did not produce significant improvement in fit, $Q(1\ df) = 1.98$, $p = .159$. However, mania scale content made a significant incremental contribution after controlling for any of the other moderators (including in the fully augmented model, below). Because this was a hypothesized moderator, we retained it in subsequent analyses.

Parent-only diagnostic interview. Another candidate moderator was whether the diagnostic interview relied solely on the parent, without the interviewer also talking directly to the youth in question. This occurred in six of the 27 samples, all of which only reported effect sizes using caregiver-rated scales. Consistent with expectations about shared source variance inflating the predictor-criterion association, interviewing only the parent produced significantly higher *g* estimates, $b = 0.62$, $Q(1\ df) = 5.80$, $p = .016$.

Distilled versus clinically representative samples. The final moderator of interest focused on the impact of sampling design. A dummy code contrasting distilled versus clinically generalizable designs accounted for significant variance in the effect sizes. Entered as the sole moderator in a multivariate metaregression, it earned a $Q(1\ df)$ of 6.64, $p = .010$, with distilled samples averaging *g* values .50 higher than the nondistilled, more generalizable samples.

Fully augmented model. A fully augmented model included all the moderators of interest simultaneously. This model accounted for substantial variance, $Q(5\ df) = 84.03$, $p < .00005$. It also reduced the random effect variance components both at Level 1 (within samples)—sigma^2 = .03 versus .13 for the model with no moderators, as well as Level 2 (between samples)—sigma^2 = .12 versus .20 in the initial model (see Table 3). There still was significant remaining heterogeneity, Cochran's $Q(57\ df) = 279.55$, $p < .00005$. The profile of likelihood plots indicated that the model provided accurate estimates, and the intraclass correlation between the estimated and true effects was 0.22 (Konstantopoulos, 2011).

Table 4 presents the regression weights and confidence intervals for the fully augmented model. The intercept was $b = 0.72$, $p < .00005$, meaning that the average effect size for caregiver report from a nondistilled, generalizable sample, using a measure that did not include specific manic symptom content, and interviewing both the youth and the caregiver, would have a $g \sim .7$. All moderators remained significant in the augmented model. Teacher report was associated with significantly lower effect sizes, $b = -0.61$, $p < .00005$, as was youth report, $b = -0.46$, $p < .00005$. Scales with specific mania item content generated moderately larger effect sizes, $b = 0.28$, $p = .004$. Relying only on the caregiver during the diagnostic interview significantly inflated effect sizes, $b = 0.42$, $p =$

Table 4
Multivariate Meta-Regression Estimates of the Effects of Moderators Entered Together in the Model

Variable	*b*	*SE*	95% CI
Intercept	.72***	.14	(.44 to .99)
Youth report (vs. caregiver)	−.46***	.09	(−.63 to −.29)
Teacher report (vs. caregiver)	−.61***	.09	(−.79 to −.42)
Mania scale content	.28**	.10	(.09 to .47)
Only parent interviewed	.42*	.21	(.01 to .83)
Distilled design	.54**	.17	(.20 to .87)

Note. CI = confidence interval.
* $p < .05$. ** $p < .005$. *** $p < .0005$, two-tailed.

.002. As hypothesized, the use of distilled samples produced much larger effect sizes than generalizable samples, $b = 0.54$, $p = .002$. Figure 2 shows the distribution of effect sizes broken down by informant and also whether or not the sample used a distilled design, graphically illustrating the effects of the two most potent moderators.

Testing the Robustness of the Meta-Regression Models

Outlier analyses. Preliminary analyses identified one study as a highly influential outlier. Reexamining the article found that the authors had reported a *SE* as if it were a *SD* (with the *T* score *SD* being less than 3, instead of close to 10). Correcting this and recalculating the effect size, the study no longer was an outlier. We corrected this before running the models reported above.

Standardized residuals flagged two studies as potential outliers in the multivariate analyses: Marchand et al. (2005) reported an effect size 1.04 *g* units larger than would be predicted based on the metaregression model, $z = 2.60$, $p < .01$. Papachristou et al. (2013) reported an effect size $g = 1.21$ units smaller than predicted based on the model, $z = -3.35$, $p < .01$. Marchand et al. (2005) used the YMRS, which has shown highly variable results across other samples. The article did not explicitly report whether the diagnoses were blind to the scale, and details about the diagnostic procedures also were sparse. Papachristou et al. used the CBCL. Rerunning the model with those two studies excluded did not change the substantive pattern of findings; all moderators remained significant with similar coefficient sizes (results available upon request from author). Controlling for year of publication, comorbidity, sample size, and a variety of other parameters that meta-analysis guidelines (American Psychological Association, 2008; Liberati et al., 2009) recommend testing also did not alter the significance or substantive pattern of results.

Publication bias. All of the analyses described above checked for influential outliers, and examined the effects of omitting outliers on sensitivity analyses. We also used the random effects, mixed model extension of Egger's regression test of publication bias (Viechtbauer, 2010a). There was no evidence of publication bias, $p > .81$ in the fully augmented model, and $p > .50$ in the model with all moderators also including ratings of reporting and design quality.

Because the MARS reporting guidelines ask for fail-safe *N,* we estimated it using two methods, separately for each informant to reduce the effects of nesting within sample. Table 5 reports the results. By every method, it appears highly unlikely that publication bias threatens conclusions about the validity of the effect sizes. In summary, all three informants produced statistically significant differences between the score distributions for bipolar versus comparison groups, with teacher report producing a medium effect size in Cohen's (1988) rubric, youth report yielding a large effect size, and caregiver report an effect size more than 50% larger than what conventionally is considered "large."

Other sensitivity analyses. Neither design quality (as operationalized by the Kowatch scoring) nor the reporting quality (operationalized as QUADAS-2 Total) moderated the observed effect sizes, either in isolation or after controlling for the other moderators (all $p > .20$); nor did including them change the significance of any of the other moderators. We also checked whether the number of scale items, the year of publication, the percentage of cases with ADHD in the sample, or whether the study had sponsorship from a pharmaceutical company had any association with the observed effect; none did after controlling for the a prior hypothesized moderators or by itself.

Exploratory Comparison of Specific Measures

We ran an exploratory version of the multivariate regression model to see if the literature supported any generalizations about the relative performance of measures. We used the effect size based on the ASEBA Externalizing score (Achenbach & Rescorla, 2001) as the comparator in a set of dummy codes that tested all scales contributing at least two independent effect sizes: the Child and Adolescent Symptom Inventory (CASI; Gadow & Sprafkin, 1994), the Child Bipolar Questionnaire (CBQ; Papolos et al., 2006), the Conners (1999), the General Behavior Inventory (GBI; Depue et al., 1981), the Mood Disorder Questionnaire (MDQ; Hirschfeld et al., 2000), and the rating scale version of the Young Mania Rating Scale (YMRS; Gracious et al., 2002). We chose the ASEBA Externalizing score as the comparator because (a) the ASEBA is the most studied scale in this review, (b) the ASEBA contributed the most effect sizes across levels of the informant and distilled moderator variables, (c) the Externalizing scale is easier for clinicians to use than other putative "bipolar" profiles, because it is a standard part of the ASEBA scoring algorithm, (d) Externalizing is not subject to concerns about overfitting to a particular sample, whereas putative "bipolar" profiles necessarily were developed post hoc on the initial sample, and (e) in all samples reporting effect sizes based on both Externalizing and multiscale profiles, the effect size based on Externalizing was larger in all but two cases. The Externalizing score essentially is an "incumbent" measure that any challenger would need to defeat to supplant it in research or practice. We did not include the dummy code for whether or not the scale content specific manic item content, because it would have been highly collinear with the scale dummy codes.

This model accounted for substantial variance, $Q(11\ df) = 111.17$, $p < .00005$. It also reduced the random effect variance components both at Level 1 – sigma^2 = .02 versus .13 for the model with no moderators, as well as Level 2 – sigma^2 = .14 versus .20 in the initial model. There still was significant remaining heterogeneity, Cochran's $Q(51\ df) = 253.85$, $p < .00005$.

The model intercept was $b = 0.72$, 95% CI [.44, 1.00]; it was the estimated true effect size for the caregiver CBCL Externalizing score from a nondistilled, generalizable sample, including both the youth and caregiver in the diagnostic interview. All four moderators remained significant in the augmented model even after controlling for scales used. Adjusting for all variables, teacher report was associated with significantly lower effect sizes, $b = -0.60$, 95% CI [−.77, −.44], $p < .00005$, as was youth report when compared to caregiver report, $b = -0.48$, 95% CI [−.63, −.33], $p < .00005$. Three scales demonstrated significant differences compared to the ASEBA Externalizing score: the MDQ averaged $g = 0.39$ higher than the Externalizing effect size, $p = .003$. The CMRS averaged $g = 0.33$ higher after controlling for all variables, $p = .024$. The GBI averaged $g = 0.31$ higher, $p = .002$. See Table 6 for full model.

Table 5
File Drawer Estimates for Disaggregated Effect Size Estimates Grouped by Informant

Effect size source	Mean *g*	Rosenberg	Orwin ($g = .20$)
Full set ($N = 63$ effects)	.80	29,819	205
Caregiver	1.06	19,087	173
Youth	.49	579	22
Teacher	.26	70	10

Note. The Rosenberg (2005) method estimates the number of unpublished studies with null findings that would be necessary to reduce the average effect size to nonsignificance, that is, not able to reject the null hypothesis that overall $g = 0$ at $p < .05$. The Orwin (1983) method estimates the number of unpublished studies with null results needed to reduce the average effect size to an a priori target magnitude. We selected $g = .20$ as the target because it is generally considered a "small" effect size and would produce negligible performance in diagnostic applications (corresponding area under the curve = .56, "poor").

Table 6
Exploratory Multivariate Meta-Regression Estimates Comparing Measures to ASEBA Externalizing Score, Controlling for Moderators

Variable	*b*	*SE*	95% confidence interval
Intercept (ASEBA externalizing)	.72***	.14	(.44 to 1.00)
CASI	.15n.s.	.22	(−.28 to .57)
CBQ	.28n.s.	.34	(−.39 to .95)
CMRS	.33*	.15	(.04 to .62)
Conners	.07n.s.	.16	(−.24 to .39)
GBI	.31***	.10	(.11 to .51)
MDQ	.39**	.13	(.13 to .65)
YMRS	.14n.s.	.11	(−.08 to .36)
Youth report (vs. caregiver)	−.48***	.08	(−.63 to −.33)
Teacher report (vs. caregiver)	−.60***	.08	(−.77 to −.44)
Only parent interviewed	.45*	.22	(.01 to .89)
Distilled design	.52**	.18	(.17 to .87)

Note. ASEBA = Achenbach System of Empirically Based Assessment; GBI = General Behavior Inventory; MDQ = Mood Disorders Questionnaire; YMRS = Young Mania Rating Scale; CMRS = Child Mania Rating Scale; CASI = Child and Adolescent Symptom Inventory; CBQ = Child Bipolar Questionnaire.
* $p < .05$. ** $p < .005$. *** $p < .0005$.

Clinical Interpretability

To provide more clinically meaningful description of results, we saved the predicted values from the metaregression, and then converted the *g* into an estimated receiver operating characteristic (ROC) AUC (using formula #4 from Hasselblad & Hedges, 1995). We also report the predicted sensitivity that could be expected for a threshold chosen to have specificity = .90 (Hasselblad & Hedges, 1995, formula #13), along with the corresponding diagnostic likelihood ratio for scoring above the specificity = .90 threshold. We report the upper and lower bounds based on the confidence intervals of the mixed model regressions. See Table 7. Figure 3 shows the estimated ROC curves for caregiver, teacher, and youth report for clinically generalizable (nondistilled) designs. The figure also includes three reference curves as benchmarks. The diagonal line represents chance performance (AUC = .50). If the base rate of bipolar were 10%, then just randomly diagnosing cases "betting the base rate" would get 10% of the bipolar cases correct (sensitivity = 10%), and 90% of the nonbipolar cases correct (specificity = 90%).

The shaded gray space marks the performance of clinical diagnosis as usual, with the accuracy of clinical diagnoses of bipolar disorder pegged at $\kappa \sim .1$ based on the converging estimates from both meta-analysis (Rettew et al., 2009) as well as recently published data (Jensen-Doss et al., 2014). Combining the κ with the estimated base rate of the target disorder allows estimation of the "diagnosis receiver operating characteristic curve" (Kraemer, 1992). With a base rate of 10%—consistent with reports from many outpatient settings—clinical diagnoses would deliver sensitivity of ~ .19, specificity of ~ .91, and an AUC of .55. The sensitivity almost doubles chance performance if clinicians were "betting the base rate," but it still is far from adequate.

Because studies do not have an objective gold standard, and even semistructured diagnostic interviews are not perfect, the reliability of the criterion diagnosis also creates a ceiling that limits test performance (Kraemer, 1992; Zhou et al., 2002). If a "perfect" predictor of bipolar disorder existed, it would still appear to be "wrong" when it disagreed with the results of the (imperfect) semistructured interview. Using κ of .80, close to the nominal rate of interrater reliability reported in those studies that included reliability information, the diagnosis curve for semistructured interviews would yield sensitivity of .82, and specificity of .98, with an AUC of .90.

Discussion

The goal of the present study was to meta-analyze the effect sizes for scales used to distinguish youth with pediatric bipolar disorder from other youth. The topic also provided an opportunity to investigate potential moderators of broad conceptual interest. These include informant effects, such as seeing if there were significant differences between youth or teacher report as compared with the responses of the

Table 7
Summary of Estimated Effect Sizes by Informant, Sample Design Type, and Mania Scale Content, Converting Hedge's g Into Diagnostic Efficiency Statistics

Sample type	Mania scale	*g*	(95% CI)	AUC	(95% CI)	Sensitivity for specificity = .90	DLR+
Caregiver report							
Generalizable	Yes	1.00	(.75 to 1.25)	.76	(.70 to .81)	.40	4.0
Generalizable	No	.72	(.44 to .99)	.69	(.62 to .76)	.29	2.9
Distilled	Yes	1.53	(1.25 to 1.82)	.86	(.81 to .90)	.64	6.4
Distilled	No	1.25	(1.01 to 1.50)	.81	(.76 to .86)	.52	5.2
Youth report							
Generalizable	Yes	.54	(.26 to .81)	.65	(.49 to .63)	.23	2.3
Generalizable	No	.26	(−.04 to .55)	.57	(.49 to .65)	.15	1.5
Distilled	Yes	1.07	(.75 to 1.39)	.78	(.70 to .84)	.44	4.4
Distilled	No	.79	(.51 to 1.08)	.71	(.64 to .78)	.32	3.2
Teacher report							
Generalizable	Yes	.39	(.10 to .69)	.61	(.53 to .69)	.18	1.8
Generalizable	No	.11	(−.19 to .41)	.53	(.45 to .62)	.12	1.2
Distilled	Yes	.93	(.59 to 1.26)	.74	(.66 to .81)	.37	3.7
Distilled	No	.65	(.36 to .94)	.68	(.60 to .75)	.26	2.6

Note. CI = confidence interval; AUC = Area Under Curve from receiver operating characteristic analysis; estimate assumes parametric distribution. Sensitivity for specificity = .90 uses same assumptions. DLR+ is the diagnostic likelihood ratio associated with scoring above the threshold attached to a specificity of .90; note that this might not be the most discriminating region of performance on a given test. All results based on design where both caregiver and youth were directly interviewed. Interviewing caregiver only would add ~.5 to the *g* estimate for all models.

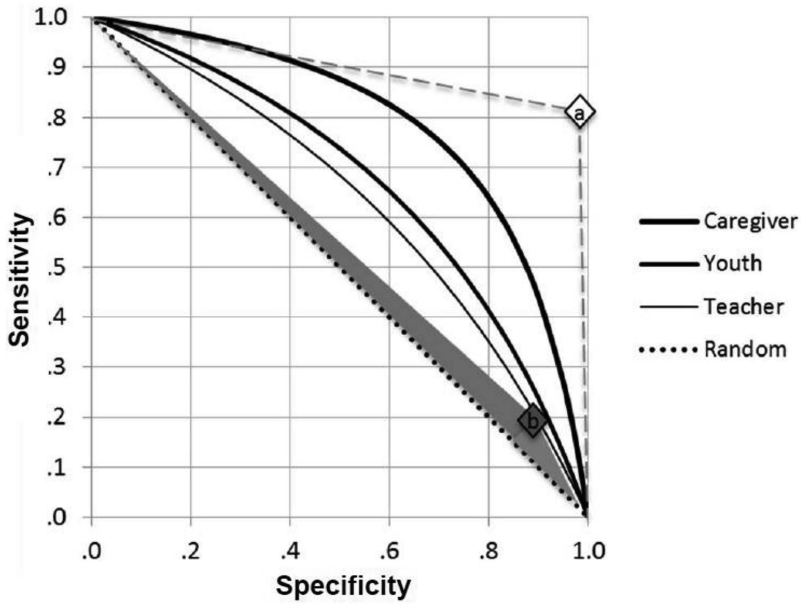


Figure 3. Plot of estimated receiver operating characteristic curves for clinical diagnosis as usual, teacher report, youth report, caregiver report, and semistructured diagnosis under clinically generalizable (nondistilled) conditions, using direct interview of both caregiver and youth to make the criterion diagnosis. *Note.* (a) denotes performance for $\kappa \sim .80$ (reported in most articles for Kiddie Schedule for Affective Disorders and Schizophrenia [KSADS] reliability) and 10% base rate; (b) denotes performance for $\kappa \sim .1$ (per Jensen-Doss et al., 2014; Rettew et al., 2009) and 10% base rate; both estimates use formula from Kraemer (1992). The area under the curve (AUC) for diagnosis as usual is .55 (shaded area), and the AUC for the semistructured interview is .90. The AUC for teacher is .61, youth is .65, and caregiver is .76.

primary caregiver for the youth (Carlson & Klein, 2014; Leibenluft, 2011; Youngstrom et al., 2006b). We also investigated design features such as whether or not the scale included specific symptoms of mania (Freeman et al., 2011; Geller & Tillman, 2005; Geller et al., 1998), whether the diagnostic interview only included the parent versus directly interviewing parent and youth, and also the effects of sampling design on the observed effect sizes (Youngstrom et al., 2006a). These design issues also apply more broadly to investigations of psychological assessment (De Los Reyes & Kazdin, 2005) and diagnostic efficiency in general (Bossuyt et al., 2003b; Zhou et al., 2002), not only in the context of pediatric bipolar disorder.

The scales in general were statistically valid and produced medium to very large effect sizes, depending upon the informant and design features. Our analyses directly modeled the multivariate, multilevel structure of the data, with many effect sizes being nested within the same sample. There was a substantial random effect variance component because of differences at the study sample level, and this variance component was consistently larger than the random effect because of variation within studies. The between-studies variance component shrank when models included our moderators of interest, but significant between-study variation remained in all models tested.

Moderators of Diagnostic Accuracy Effect Sizes

Informant effects. Caregiver report produced substantially larger effect sizes than teacher or youth report, whether the model included no other variables, all hypothesized moderators, or even during all sensitivity analyses. The gap between caregiver versus youth report was $g = -0.46$ after controlling for all other moderators. The gap between caregiver and teacher report was slightly larger: -0.61 after controlling for all other moderators. The CIs for youth and teacher report overlapped to a large extent, indicating that they performed similarly. The finding that caregiver report yielded larger effect sizes than youth or teacher report is consistent with the few prior studies that directly tested differences between the accuracy of informants (Youngstrom, Findling, Calabrese, et al., 2004; Youngstrom et al., 2005). Among all the samples that contained effect sizes based on both caregiver and youth report, the caregiver effect size was larger in every instance except for the sample from South Korea (Lee et al., 2014). Similarly, in every sample providing both caregiver and teacher rating effect sizes, the caregiver report was larger.

The high validity for caregiver report is reassuring in many ways. Caregivers are more likely to initiate seeking services for the child, rather than the child or adolescent self-referring, and interviewers tend to perceive caregivers as a more credible source of information about the youth's functioning, especially for younger children (Youngstrom et al., 2011). Caregivers also are a primary source of information about developmental history, family mental health history, and other factors crucial to the diagnostic and evaluation process (Richters, 1992). The greater validity for caregiver report also makes sense; caregivers are likely to have better reading ability and be more psychologically minded than young children or adolescents. Plus, caregivers notice symptoms such as irritable mood at lower levels of mania, whereas the mania needs to be markedly more severe before youths endorse the same symptom (Freeman et al., 2011). Many other symptoms of hypomania and mania also are likely to distress other people before they bother the person expressing the symptom.

Youth report also consistently produced statistically significant differences between bipolar and nonbipolar comparison groups, but the effect size would conventionally be considered "moderate," and it translates into modest performance in terms of diagnostic efficiency. Some general reasons that youth report might show lower validity include factors undermining the reliability of youth report, such as

poorer reading ability or lack of motivation to complete scales thoughtfully (Sattler, 2002). The content of items might be less developmentally appropriate, although that is unlikely to be a factor for instruments designed specifically for use with children and adolescents, such as the Achenbach System for Empirically Based Assessment or newer mania scales such as the CBQ and CMRS. In the context of evaluating potential bipolar disorder, youth report may show diminished validity to the extent that loss of insight into one's behavior is a feature of hypomania and mania (Youngstrom, Findling, & Calabrese, 2004), as has been observed with adults with bipolar disorder or psychosis (Dell'Osso et al., 2002). A practical consideration is that reading level precludes using most of these scales with youths younger than age 11 years, and normative data for instruments such as the ASEBA and Conners do not extend below age 11, either.

Teacher report produced effect sizes substantially smaller than caregiver report, and somewhat lower than youth report. Though teacher report still produced statistically significant differences between the bipolar and comparison group, the effect size was small to medium. Translated into an area under the curve, teacher report produced AUC values of ~.60 for mania scales in nondistilled samples, often considered the "poor" range for clinical utility (Swets, Dawes, & Monahan, 2000). Some of the symptoms more specific to mania, such as a decreased need for sleep, are difficult to observe in a classroom setting. Many behaviors readily seen in the classroom are also symptoms that overlap with ADHD, reducing the diagnostic specificity of teacher report.

Overall, the differences in performance across informant are consistent with the patterns that have emerged in the few head-to-head comparisons within the same sample (Geller et al., 1998; Hazell et al., 1999; Youngstrom, Findling, Calabrese, et al., 2004; Youngstrom et al., 2005). The correlation between caregivers, youths, and teachers about the youth's behavior problems tends to be modest, and some studies have found that the degree of problems reported by teachers or youths is actually significantly higher in cases with bipolar disorder than would typically be predicted based on caregiver report (Youngstrom et al., 2006b). However, the general level of problems endorsed by teachers or youths is much lower than by caregivers across almost all of the samples (with the exception of the data from South Korea). It is possible that a combination of selection bias and regression artifacts is contributing to the results: because most of the effect sizes come from clinical outpatient samples, the caregiver concerns were the driving factor for the bulk of referrals. Effectively, outpatient samples select participants on the basis of high levels of caregiver reported problems. Because caregiver and youth or teacher report usually correlate less than $r \sim .3$, regression to the mean dictates that the average youth or teacher scores will be much lower. The attenuation will shrink the scores for the bipolar group more, inasmuch as the caregiver reported higher concerns for them, thus reducing the effect sizes comparing the bipolar to other groups.

It also is possible that youth behavior changes between school and home settings. Some argue that mania should be pervasive and observable across multiple settings (Carlson, 2011). Certainly severely disorganized behavior would be easy to note, but other symptoms, such as irritability, difficulty concentrating, or energy changes may not be remarkable in a classroom setting. Additionally, there are likely to be circadian fluctuations in mood and energy (Murray et al., 2009). To the extent that bipolar disorder is associated with an evening chronotype or delay of sleep phase, the attendant mood changes are likely to be most pronounced outside of school hours (Harvey, 2008). The emotional significance of relationships also may contribute to differences in behavior. If rejection sensitivity and a sense of intimacy are salient issues for youths with bipolar disorder (Ehnvall et al., 2011; Robertson et al., 1996), then the home life is more emotionally significant than school, at least until peers become more important. A third process that could lead to more conflict at home is the testing of developmental boundaries and the youth's push for greater control and autonomy (Emery, 1992). At present, there are no published studies using direct observation of interactions at home involving youths with bipolar disorder. Actimetry and other objective methods for measuring youth behavior could be informative about the extent to which behavior changes between settings or dyads (Axelson et al., 2003).

The consistent finding that caregiver report produces larger effect sizes than youth or teacher report indicates that the revisions to the psychiatric nosology were justified in not requiring elevated teacher report of manic symptoms to confirm a diagnosis of pediatric hypomania or mania (American Psychiatric Association, 2013; Youngstrom et al., 2008). The validity of parent reported mood symptoms also is supported by sensitivity to treatment effects in double-blinded clinical trials (e.g., Findling et al., 2007; Findling, McNamara et al., 2005; West, Celio, Henry, & Pavuluri, 2011), as well as brain imaging studies—where parent report produces larger correlations with patterns of activation in the youth's brain than diagnostic categories or youth ratings do (e.g., Bebko et al., 2014).

Mania scale content. Results also supported our hypothesis that scales including symptoms more specific to mania would show larger effect sizes. The ASEBA and Conners were written at a time when mania was considered an "adult only" phenomenon, and so many relevant items were not included in the pool. Both include manic symptoms, but they are ones that are not diagnostically specific and, thus, often attributable to other conditions, as indicated by the factor structures in the measures' normative samples. More recent versions of some broad checklists, such as the BASC and the CASI, have added a mania scale to at least some informant versions. Thus they may perform more similarly to instruments such as the MDQ, CMRS, and CASI that inquire directly about *DSM* symptoms, or the GBI and CBQ, which also cover associated features beyond the core *DSM* symptoms.

The source articles did not report sufficient statistics for us to examine directly whether the inclusion of mania symptoms improved the specificity more than the sensitivity of the scales. However, the greater effect on specificity is plausible because the more sensitive symptoms of bipolar disorder tend to be nonspecific features such as irritable mood and difficulty concentrating. These symptoms are well represented on both general, broad-coverage instruments as well as mania-oriented scales. In contrast, more specific symptoms such as decreased need for sleep and elated mood tend to only be included on scales purpose-built to investigate manic symptoms.

With the exception of the CASI, the commercially distributed broad-coverage instruments do not have a mania scale containing diagnostically specific symptoms, or there are not yet published data indicating the diagnostic performance of any new scale that they have added in the most recent revisions. The CBCL Externalizing score, or the putative bipolar profiles of subscales, show good sensitivity but poorer specificity. This positions the broad measures to be good at ruling out cases of bipolar disorder, but makes them prone to high false positive rates if used alone to screen for bipolar (Straus et al., 2011). Positive results on tests with moderate specificity are ambiguous because there is a high false alarm rate (the other side of the low specificity coin). Combined with the base rate of bipolar being low in most settings, the result is low positive predictive values. Put bluntly, most cases "testing positive" for bipolar disorder on measures that do not focus on diagnostically specific symptoms will not actually have

bipolar disorder unless the setting is an inpatient unit or similar venue where the base rate of bipolar is high.

Interview strategy: The importance of laying eyes on the child. Relying solely on the primary caregiver during the semistructured interview also changed the average effect size significantly. Six of the 27 samples relied solely on the caregiver for the diagnostic interview. Consistent with concerns about shared source variance (Carlson & Klein, 2014; Podsakoff, MacKenzie, & Podsakoff, 2012), relying only on one person's perspective inflated the association between the caregiver rating and the diagnosis, increasing the observed effect size by g ~.6. The six samples only reported effect sizes for caregiver-reported scales, where the influence of shared source variance would be largest. It is interesting that these six studies ran the gamut in terms of youth ages, with the average age extending from 8.7 years (Biederman et al., 1995) to 16.0 years (Papachristou et al., 2013), so results were not driven by youth age. Findings reinforce the value of directly interviewing the youth, even when they may be too young to complete a full semistructured diagnostic interview. In young children, direct observation during the interview may provide an opportunity to observe behaviors that might indicate the presence of a pervasive developmental disorder or other alternate explanations for child behavior (Carlson & Klein, 2014). The credibility and sophistication of youth perspective increases steadily with age and verbal ability (Youngstrom et al., 2011), making the older youth's input more valuable. Integrating multiple sources of information is likely to combat factors that would undermine the validity of any single informant, such as demoralization, malingering, or attempting to minimize problems (Spitzer, 1983). Inasmuch as diagnoses that synthesize multiple information sources are likely to be more valid, estimates of diagnostic accuracy that are pegged to such diagnoses are likely also to have better validity, even if the size of the coefficient itself appears more humble.

Distilled sample enrollment. The final moderator variable of interest was the design of the sample. Whereas diagnostic sensitivity and specificity were once thought to be intrinsic properties of a measure, methodologists now realize that these parameters can change markedly as a result of design features. The present findings show that this is not an abstract concern. Differences in sampling design changed the observed effect sizes by $g \sim .56$, even after controlling for other moderators. This ~.5 bias is consistent with the findings of prior work that compared the effects of distilled versus more generalizable sampling inclusion criteria in the same data set (Youngstrom et al., 2006a). Taking the eight scales examined in that study and comparing the effect sizes observed in distilled versus nondistilled designs (reported in Table 4 of Youngstrom et al., 2006a) found an average upward bias of $g = +.43$. The PBD literature includes many studies focused on phenomenology and careful descriptive validation of the syndrome. These studies often included healthy control comparison groups, and they also often had stringent exclusion criteria that reduced the amount and types of clinical heterogeneity observed. Although these designs were internally valid for their intended purpose, they are less generalizable to typical clinical practice. The results of the meta-analysis underscore that the reduced generalizability produces systematic bias that exaggerates the diagnostic efficiency of scales. Distilled samples create a rising tide that lifts all boats, inflating the effect sizes of all scales and potentially boosting less valid tests even more than others (Youngstrom et al., 2006a). This is especially true for scales with nonspecific items, such as externalizing or attention problems, which also would be elevated in groups with other diagnoses but not in healthy controls (Yeh & Weisz, 2001).

The positive bias is pernicious for two reasons. First, it obscures differences between scales. The boost from a distilled design is larger than the difference between most of the valid measures. Design effects swamp the relative differences between scales, potentially making weak scales look better than a more valid scale tested under more generalizable conditions. Second, the exaggerated effect sizes in a distilled design will bias the clinical interpretation of the scales. Inflated sensitivity and specificity estimates lead to more extreme diagnostic likelihood ratios, and more extreme predictive values. If clinicians rely on a simplistic "positive test result" interpretation, the false positive rate will be higher than it appears. Bad designs, in terms of validity for studying diagnostic efficiency, will contribute to overdiagnosis of bipolar disorder. More rigorous and generalizable designs produce more humble estimates of effect size, but these are essentially "preshrunk" to fit their application across a broader range of clinical settings.

The clinical "take home" messages from this moderator analysis reinforce the dicta from Evidence Based Medicine to consider both the validity of the study design, and also whether the participants look like the patients with whom the clinician is working. As receiver operating characteristic analyses become more popular, test consumers should cultivate healthy skepticism: if a test claims to produce excellent results at a challenging task, check carefully to see if design flaws are swelling the effect size. It also behooves researchers to consider the potential biasing effects of design features ahead of time before conducting secondary analyses of data originally gathered for a different purpose than studying diagnostic efficiency. The STARD Guidelines and QUADAS-2 checklist were intended in part to provide a convenient checklist for this sort of rapid evaluation of reports.

Exploratory Comparison of Different Scales

We also ran exploratory analyses comparing the performance of measures, adjusting for the moderator variables, and accounting for correlated effects because of nesting within sample. These analyses should be considered tentative, as the available literature has gaps in coverage. Furthermore, the mixed model metaregression would have less statistical power for direct comparison of measures than would be possible by applying best methods directly to the raw data. However, even with these caveats, the analyses indicated that the GBI, MDQ, and CMRS all performed significantly better than the ASEBA, reflecting that their item sets include the symptoms more specific to bipolar disorder. Reassuringly, the results of the meta-analysis align with the results of the head-to-head comparisons that have been done using raw data in the past. The CASI is likely to also perform better than the ASEBA or other nonspecific measures, based on item content and observed effect size, but the critical region around the observed effect is large based on the analysis only having two effect sizes for the CASI.

Reporting Quality and Sensitivity Analyses

Reporting quality. Assessed against the criteria developed by Kowatch et al. (2005), the design of the studies tended to be good. The samples included large numbers of cases, used semistructured diagnostic interviews, and routinely assessed common comorbidities. The quality of reporting of results was less strong. The bulk of the publications predated the dissemination of the STARD and other reporting guidelines, so it is not surprising that there were some omissions. Key places for improvement include adding flow diagrams, documenting the length of time between checklists and diagnoses, and clarifying whether all participants were included in the analyses. The bad news is that there was no significant trend for the reporting quality to improve in more recent studies, though hopefully that will change. The good news is that reporting quality was not

related to the effect sizes, nor did any of the tests of moderators change when adjusting for quality. This does not mean that reporting quality is unimportant, merely that variations in quality did not add significant bias to the observed effect sizes.

Consideration of Alternative Explanations

A standard concern with meta-analysis is whether the published literature diverges from results that were never published—the "file drawer problem." The literature reviewed here is less prone to publication bias, for two reasons. Many of the studies included in the review were descriptive, phenomenological studies or epidemiological studies. For these articles, statistical significance or rejection of a null hypothesis was not a major consideration for "publishability," making it unlikely that nonsignificant results would be censored from the published literature. The second reason is that measures used for screening and diagnosis require large effect sizes to achieve respectable accuracy rates for classification (Hummel, 1999; Youngstrom & De Los Reyes, 2015). Consequently, statistical significance is an easy bar to surpass, even with relatively small sample sizes. Consistent with these scenarios, Egger's test found no evidence of publication bias in the multivariate models. We also ran file drawer analyses separately for caregiver, youth, and teacher report. This is conservative, because it split the overall sample into pieces with fewer effect sizes and constituent cases. Teacher report offered the worst-case scenario, as it included the fewest effect sizes, the least participants, and the smallest average effect size. Even for it, the file drawer would need to be loaded with 10 null studies before the mean effect size for teacher report would decrease to a "small" effect size ($g \sim .2$), and 70 null studies before teacher report would no longer be significant $p < .05$. For caregiver report, the number of unpublished studies would need to be more than 19,087 to push the p value greater than .05. Publication bias does not seem to be a major threat here.

A conceptually more interesting issue is potential circularity between the information source for the scale and the criterion diagnosis. This is not the same thing as "criterion contamination," where the diagnostician would have access to the scale scores when formulating the diagnostic impression. The older concept of shared method variance is closer to the mark (Campbell & Fiske, 1959). If the same informant fills out two rating scales about different constructs, the resulting scores will correlate with each other because of shared method variance (Podsakoff et al., 2012). For the studies reviewed in the meta-analysis, one of three different informants completed the rating scales, and a subset of the same informants contributed to the diagnostic interview. In the most extreme scenario, the caregiver might be the only person interviewed about the youth's diagnosis, and she or he also completed the rating scale. On one hand, the interview still involves additional information beyond that gleaned from a scale—there are opportunities for probing, follow-up questions, interpretation of nonverbal cues, and clinical judgment (Garb, 1998; Kaufman et al., 1997). On the other hand, caregiver reported scales might have an unfair advantage if the caregiver's perspective heavily influences the diagnostic interviews. The apparent advantage for caregiver report in terms of larger effect size could be due in part to the shared source variance contributing to both predictor and criterion diagnosis.

This concern is allayed somewhat by the fact that caregiver report showed large effects across all studies, regardless of whether the other informants participated in the diagnostic interview. Parent report's validity also is supported by sensitivity to treatment effects in double-blinded clinical trials (e.g., Findling et al., 2007; Findling, McNamara et al., 2005; West et al., 2011): the double-blind design controls for expectancy effects, as they would contribute to perceived improvement in the placebo arm (Keck, Welge, Strakowski, Arnold, & McElroy, 2000). Furthermore, in brain imaging studies parent report produces larger correlations with patterns of activation in the youth's brain than diagnostic categories or youth ratings do (e.g., Bebko et al., 2014). It also is also noteworthy that the source variance artifact would inflate agreement with caregiver report across all diagnoses. However, there are published examples where youth report shows the same validity estimates as caregiver report for anxiety diagnoses (Van Meter et al., 2014) or significantly higher validity for reports of posttraumatic stress disorder (You, Youngstrom, Feeny, Youngstrom, & Findling, 2015). These are secondary analyses of some of the same data used in the meta-analysis here and found larger effects for caregiver report predicting bipolar diagnoses (Youngstrom, Findling, Calabrese, et al., 2004; Youngstrom et al., 2005). This indicates that the validity of the caregiver report is higher for bipolar disorder than for some other diagnoses, showing a degree of specificity that argues against a general methodological artifact. Converging evidence also suggests that manic symptoms are associated with a loss of insight that undermines self report, and many symptoms are likely to be noted earlier by collateral informants (Freeman et al., 2011)—although the weaker performance of teacher report still needs to be reconciled with the higher validity of caregiver report in this regard. Situational specificity in behavior, and dyadic patterns of interaction, remain key topics for further exploration.

Generalizability of Conclusions

The literature on the assessment of pediatric bipolar disorder has grown rapidly. At this point it spans a range of measures, multiple research groups, seven different countries drawn from four continents, and five languages of administration. Samples came mostly from outpatient settings, with some high risk offspring samples and an epidemiological sample also contributing. Within the limited range represented in the meta-analysis, clinical setting appears less important than diagnostic inclusion and exclusion criteria.

As these measures are translated and used with more diverse populations, it will be important to test whether the psychometrics change across demographic groups. The few studies done in this regard suggest that the measurement properties of depression and hypomania/mania tend to be robust across race and ethnicity (Gamma et al., 2013; Pendergast et al., 2015), and the measures show consistent diagnostic efficiency across race/ethnicity within the United States (Jenkins et al., 2012). This is reassuring, and suggests that the measures are likely to help improve assessment across a range of demographic groups. In the case of bipolar disorder, they could directly contribute to reducing race differences in diagnoses, where minority groups with mood disorder have been particularly vulnerable to misdiagnosis in the United States.

The data from Korea stand as the notable exception to this overall pattern. It is a single sample, but, it is intriguing that the Korean study used two of the most valid rating scales (the MDQ and the GBI) and found that youth report exceeded parent report on both. Cultural factors, including high levels of stigma against mental health concerns, tremendous parental emphasis on educational excellence, perfectionism, and differences in familial patterns of communication all deserve exploration. Rapid economic and cultural change in Korea also may create age cohort effects, where the adolescents may have more Westernized attitudes toward mood and mental health, and greater awareness and acceptance of mental health issues. These hypotheses would best be addressed by a combination of qualitative studies and replications in other cultures with a high degree of Confucian values, such as Taiwan or Hong Kong, as well as countries with rapidly developing economies but different cultural traditions, such as Chile or India (Meeuwesen, van den Brink-Muinen, & Hof-

stede, 2009; Minkov & Hofstede, 2011). Present results indicate a good degree of generalizability across groups within the United States and other Westernized countries, with a big asterisk qualifying the validity of parent report in Asian cultures.

Limitations

No meta-analysis can be stronger than the literature upon which it is based. Because many of the primary articles predated the publication of recent reporting guidelines (American Psychological Association, 2008; Bossuyt et al., 2003a; Liberati et al., 2009), it is not surprising that they did not include many of the suggested elements, such as flow diagrams. The quality of design (aside from distilled samples) and the quality of reporting did not significantly moderate the results of these meta-analyses. Still, it will be helpful for the field to make a conscious effort to improve clarity of reporting about key design features. Several other factors potentially influencing the distribution of scores in cases with bipolar disorder were not reported frequently enough to be included in the meta-analysis. These included things such as the rates of bipolar I, bipolar II, cyclothymic, and NOS cases, or the current mood state of cases. As noted in the introduction, cases with more severe current presentation will on average have higher scores on the scales, making the diagnostic sensitivity higher. Conversely, the published reports also did not include enough details about the diagnostic composition of the comparison group to support investigation of effects on diagnostic specificity. These types of questions could be fruitful topics for a "mega-analysis" pooling raw data from multiple samples.

Although all three of our hypothesized moderators produced significant effects, there still remained significant heterogeneity in effect sizes both within and between samples. We tested standard candidate moderators, such as publication year, and did not identify other significant moderators. However, the heterogeneity suggests that other factors exist that were not coded or reported in the literature. Rates of comorbidity or inclusion of different cognate diagnoses—such as ADHD, depression, or conduct disorder—are a likely contender for accounting for some of this variance. Differences in rater training also may be a source of variation in the definition of the criterion measure. Semistructured interviews give greater latitude to clinical judgment, making the differences because of training potentially sizable (Dubicka, Carlson, Vail, & Harrington, 2008; Mackin, Targum, Kalali, Rom, & Young, 2006). Conversely, more structured approaches to interviewing may increase the consistency, falling along a continuum to a place where all that differs is whether the items are read by the participant or the interviewer when no alternative phrasing or clinical judgment is allowed. Taken to that extreme, the increased associations could be a form of pseudovalidity if produced by shared source variance versus greater content validity (Garb, 1998; Spitzer, 1983).

Implications for Theory, Policy, and Practice

The results of the meta-analysis confirm several important theoretical points. The first is that caregiver report produces larger effects for assessing bipolar spectrum disorder than self report or teacher report. Research studies investigating the correlates of youth mood symptoms, such as genetic or imaging studies, would do well to include caregiver-reported measures of mood symptoms.

At a policy level, these findings support the *DSM–5* decision not to require impairment in multiple settings or across multiple informants as a requirement for diagnosing bipolar disorder in youths (American Psychiatric Association, 2013; Angst, 2013; Youngstrom, 2009). Practice guidelines should (a) emphasize gathering caregiver report when possible to clarify assessment questions around potential bipolar disorder, and (b) not require elevation on report from multiple raters about manic symptoms, recognizing that cross-informant agreement tends to be low in general. Evaluation pathways for mood disorder also should incorporate caregiver-reported measures. The fact that several free and brief measures produced some of the best effect sizes in the meta-analysis improves the feasibility of adoption.

With regard to clinical practice, several measures are now well established as tools for evaluating bipolar symptoms in youth. In the treatment literature, having at least two published reports conducted by independent research groups offers protection from allegiance effects as well as quirks of an individual study (Chambless & Hollon, 1998). Using a similar criterion of two independent replications, five caregiver report measures have established clinically meaningful effect sizes: the GBI, MDQ, CMRS, ASEBA, and CBQ. The CASI and Conners scales have been evaluated in one sample each, and the YMRS performance is too varied and poor to endorse when so many better alternatives are available. The MDQ and GBI also appear adequate as a self-report measure.

How do the checklists compare to the other alternatives available for assessing potential pediatric bipolar disorder? Benchmarking the effect sizes from this meta-analysis against those reported in other reviews shows that caregiver report on best measures yields effect sizes larger than generated by neurocognitive tasks comparing cases with bipolar disorder to ADHD or healthy controls (Joseph, Frazier, Youngstrom, & Soares, 2008; Walshaw, Alloy, & Sabb, 2010). Recent studies have applied machine learning algorithms to functional magnetic resonance imaging (fMRI) data as a way of discriminating bipolar disorder (Rocha-Rego et al., 2014). These studies involve both overfitting, where the algorithm is optimized for the sample at hand, and "distilled designs" that produced significantly larger effect sizes in this meta-analysis. Even with both of those upward biases, the algorithm produced an AUC of .78. The fMRI methods will produce smaller effects when used under clinically realistic conditions with greater diagnostic heterogeneity, especially given the transdiagnostic nature of brain regions involved in affect regulation (Hartley & Phelps, 2010; Strakowski et al., 2012). Even when fMRI or neurocognitive algorithms evolve to produce comparable or larger effect sizes in generalizable designs, there still are issues of greater cost and more limited accessibility (Kraemer, 1992). For the foreseeable future, checklists appear to be the frontrunner in an important niche of clinical assessment despite their imperfections.

Guidelines for Future Research

Present findings limn several guidelines and priorities for future research.

Avoid artifacts—especially criterion contamination and distilled designs. Consult the STARD guidelines when designing studies of diagnostic measures (Bossuyt et al., 2003b). Although the STARD guidelines were published in multiple journals and endorsed by multiple editorial boards (e.g., Meyer, 2003), it is sobering that there is no difference in the average quality of reporting of studies published before or after the STARD guidelines were promulgated.

Similarly, researchers need to consider design issues when deciding to do ROC as a secondary analysis (Youngstrom, 2014; Zhou et al., 2002). Using biased designs is not an abstract problem—it was pandemic in the literature that we reviewed, and it produced a large bias in the observed results. Comparisons with healthy controls are clinically trivial, and make all measures look good (Youngstrom et al., 2006b). These designs produce exaggerated effect sizes, which will translate into excessive pseudoaccuracy of clinical decisions. The exclusion of competing diagnoses that are clinically common, such as unipolar depression or conduct

disorder, will reduce the number of false positive cases in the research report, biasing the apparent diagnostic specificity upward. Using the same threshold in a clinical setting that does not exclude these cognate diagnoses will lead to much higher rates of false positives. Put simply, using results from distilled designs will contribute to overdiagnosis when applied in most clinical contexts. Distilled sampling was the single most potent moderator we identified in the metaregressions. Please do not publish results if based on a biased design—these data are not helpful. Findings that have criterion contamination where the same measure is used to define the proxy diagnosis and then predict are equally problematic. They contribute noise to the literature and risk confusing consumers about choice of measure and interpretation. Peer reviewers should treat these as serious, perhaps fatal, flaws when reviewing manuscripts. On the STARD list of 25 design and reporting considerations, these are 800 pound gorillas; and they have been swinging amok through the mood assessment literature. We excluded a half-dozen studies that used proxy definitions with criterion contamination; and 52% of the samples (44% of the effect sizes) included in the meta-analysis used distilled designs, creating a gargantuan artifact in the literature.

Add objective criterion measures besides diagnosis. The next generation studies examining the validity of mood ratings should use objective, heteromethod measures to explore differences in validity of teacher, caregiver, and youth report. Benchmarking these against actimetry (Trull & Ebner-Priemer, 2013), gene expression, brain imaging of core constructs (e.g., Bebko et al., 2014), or ecological validators such as high risk behavior (e.g., Stewart et al., 2012) all will help triangulate the relative weight clinicians and policymakers need to assign each perspective. These data also will inform future nosological revisions, and they will map the connections between levels of functioning that the NIMH Research Domain Criteria (RDoC) initiative aims to organize conceptually (Cuthbert & Insel, 2010).

Explore culture as a moderator. Another goal is to prioritize understanding cultural differences contributing to performance. Within the United States, the few studies that have investigated differential item functioning or structural invariance tend to find that bipolar mood scales tend to show little bias. This contrasts sharply with the disparities in clinical diagnoses and service utilization data. The burden of depression and bipolar disorder is equally serious in the developed and developing world (Gore et al., 2011). If these assessment tools are similarly valid across cultures, they offer an inexpensive method for early identification and intervention that can be implemented on a much larger scale than semistructured interviewing or biological/neurocognitive assays (Kraemer, 1992). Part of the discrepancy between assessment and diagnostic findings could be closed by greater use of evidence based assessment methods, using the best of the rating scales and interpreting them within an EBM/Bayesian approach.

However, there also is likely to be a context of cultural differences in beliefs about the causes of emotional and behavioral issues, as well as attitudes toward treatment. The Korean data in the meta-analysis are an intriguing anomaly that suggests a potentially powerful yet nuanced role for culture. More research needs to include groups with traditionally Confucian and other non-Western views (Cheung & Leung, 1998; Meeuwesen et al., 2009), not just because it is an interesting academic question, but because most of the human population lives in these cultures, and these factors change the dynamic of medical communication between practitioner and patient (Meeuwesen et al., 2009).

Study collateral informants in adulthood and late life. Whereas including collateral informants is routine with children and adolescents, it is the exception with adults. The difference in perception of behaviors by self versus other is a robust phenomenon, including the "fundamental attribution error" in social psychology. *DSM–5*s new emphasis on energy in the mania A criterion also flowed from recognition that memory may be more accurate for changes in energy than mood (Angst, 2013). The larger effect size for caregiver report of manic symptoms deserves exploration in other age groups, especially given the loss of insight tied to hypomania and mania (Dell'Osso et al., 2002; Young, Biggs, Ziegler, & Meyer, 1978) and the tremendous interpersonal consequences of manic episodes (Algorta et al., 2011; Du Rocher Schudlich, Youngstrom, Calabrese, & Findling, 2008; Miklowitz, 2002).

Present results also emphasize the importance of prioritizing dissemination of assessment tools and teaching evidence based assessment methods. The scales that performed best in the meta-analysis are in the public domain, but they are not heavily advertised or widely taught in training programs (Camara, Nathan, & Puente, 2000; Stedman, Hatch, & Schoenfeld, 2001). The meta-analysis has established that there is a set of caregiver report measures that delivers clinically meaningful effect sizes even when evaluated under externally valid designs. These could produce large improvements in clinical decisions (Youngstrom, Choukas-Bradley, Calhoun, & Jensen-Doss, 2015), potentially even more so with underserved minority groups (Jenkins et al., 2012; Pendergast et al., 2015).

References

*References marked with an asterisk indicate studies included in the meta-analysis.

Ablow, J. C., Measelle, J. R., Kraemer, H. C., Harrington, R., Luby, J., Smider, N., . . . Kupfer, D. J. (1999). The MacArthur Three-City Outcome Study: Evaluating multi-informant measures of young children's symptomatology. *Journal of the American Academy of Child & Adolescent Psychiatry, 38,* 1580–1590. http://dx.doi.org/10.1097/00004583-199912000-00020

Achenbach, T. M. (2001). What are norms and why do we need valid ones? *Clinical Psychology: Science and Practice, 8,* 446–450. http://dx.doi.org/10.1093/clipsy.8.4.446

Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the child behavior checklist and revised child behavior profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213–232. http://dx.doi.org/10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont.

Algorta, G. P., Youngstrom, E. A., Frazier, T. W., Freeman, A. J., Youngstrom, J. K., & Findling, R. L. (2011). Suicidality in pediatric bipolar disorder: Predictor or outcome of family processes and mixed mood presentation? *Bipolar Disorders, 13,* 76–86. http://dx.doi.org/10.1111/j.1399-5618.2010.00886.x

Althoff, R. R., Ayer, L. A., Rettew, D. C., & Hudziak, J. J. (2010). Assessment of dysregulated children using the Child Behavior Checklist: A receiver operating characteristic curve analysis. *Psychological Assessment, 22,* 609–617. http://dx.doi.org/10.1037/a0019699

American Psychological Association. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851. http://dx.doi.org/10.1037/0003-066X.63.9.839

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Angst, J. (2013). Bipolar disorders in *DSM–5*: Strengths, problems and perspectives. *International Journal of Bipolar Disorders, 1,* 12. http://dx.doi.org/10.1186/2194-7511-1-12

Angst, J., Azorin, J. M., Bowden, C. L., Perugi, G., Vieta, E., Gamma, A., & Young, A. H., & the BRIDGE Study Group. (2011). Prevalence and characteristics of undiagnosed bipolar disorders in patients with a major depressive episode: The BRIDGE study. *Archives of General Psychiatry, 68,* 791–798. http://dx.doi.org/10.1001/archgenpsychiatry.2011.87

Angst, J., Gamma, A., Bowden, C. L., Azorin, J. M., Perugi, G., Vieta, E., & Young, A. H. (2012). Diagnostic criteria for bipolarity based on an inter-

national sample of 5,635 patients with *DSM–IV* major depressive episodes. *European Archives of Psychiatry and Clinical Neuroscience, 262,* 3–11. http://dx.doi.org/10.1007/s00406-011-0228-0

Anthony, J., & Scott, P. (1960). Manic-depressive psychosis in childhood. *Child Psychology and Psychiatry, 1,* 53–72. http://dx.doi.org/10.1111/j.1469-7610.1960.tb01979.x

Axelson, D. A., Bertocci, M. A., Lewin, D. S., Trubnick, L. S., Birmaher, B., Williamson, D. E., . . . Dahl, R. E. (2003). Measuring mood and complex behavior in natural environments: Use of ecological momentary assessment in pediatric affective disorders. *Journal of Child and Adolescent Psychopharmacology, 13,* 253–266. http://dx.doi.org/10.1089/104454603322572589

Axelson, D., Birmaher, B., Strober, M., Gill, M. K., Valeri, S., Chiappetta, L., . . . Keller, M. (2006). Phenomenology of children and adolescents with bipolar spectrum disorders. *Archives of General Psychiatry, 63,* 1139–1148. http://dx.doi.org/10.1001/archpsyc.63.10.1139

Axelson, D., Findling, R. L., Fristad, M. A., Kowatch, R. A., Youngstrom, E. A., Horwitz, S. M., . . . Birmaher, B. (2012). Examining the proposed disruptive mood dysregulation disorder diagnosis in children in the Longitudinal Assessment of Manic Symptoms study. *The Journal of Clinical Psychiatry, 73,* 1342–1350. http://dx.doi.org/10.4088/JCP.12m07674

Bebko, G., Bertocci, M. A., Fournier, J. C., Hinze, A. K., Bonar, L., Almeida, J. R., . . . Phillips, M. L. (2014). Parsing dimensional vs diagnostic category-related patterns of reward circuitry function in behaviorally and emotionally dysregulated youth in the Longitudinal Assessment of Manic Symptoms study. *Journal of the American Medical Association Psychiatry, 71,* 71–80. http://dx.doi.org/10.1001/jamapsychiatry.2013.2870

*Biederman, J., Faraone, S., Mick, E., Wozniak, J., Chen, L., Ouellette, C., . . . Lelon, E. (1996). Attention-deficit hyperactivity disorder and juvenile mania: An overlooked comorbidity? *Journal of the American Academy of Child & Adolescent Psychiatry, 35,* 997–1008. http://dx.doi.org/10.1097/00004583-199608000-00010

Biederman, J., Klein, R. G., Pine, D. S., & Klein, D. F. (1998). Resolved: Mania is mistaken for ADHD in prepubertal children. *Journal of the American Academy of Child & Adolescent Psychiatry, 37,* 1091–1099. http://dx.doi.org/10.1097/00004583-199810000-00020

Biederman, J., Mick, E., Faraone, S. V., Spencer, T., Wilens, T. E., & Wozniak, J. (2003). Current concepts in the validity, diagnosis and treatment of paediatric bipolar disorder. *The International Journal of Neuropsychopharmacology, 6,* 293–300. http://dx.doi.org/10.1017/S1461145703003547

*Biederman, J., Wozniak, J., Kiely, K., Ablon, S., Faraone, S., Mick, E., . . . Kraus, I. (1995). CBCL clinical scales discriminate prepubertal children with structured interview-derived diagnosis of mania from those with ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry, 34,* 464–471. http://dx.doi.org/10.1097/00004583-199504000-00013

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . de Vet, H. C. W., & the Standards for Reporting of Diagnostic Accuracy. (2003a). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326,* 41–44. http://dx.doi.org/10.1136/bmj.326.7379.41

AQ: 7

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . Lijmer, J. G., & the Standards for Reporting of Diagnostic Accuracy. (2003b). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry, 49,* 7–18. http://dx.doi.org/10.1373/49.1.7

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. http://dx.doi.org/10.1097/00004583-199207000-00006

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31,* 141–154. http://dx.doi.org/10.1037/0735-7028.31.2.141

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105. http://dx.doi.org/10.1037/h0046016

Carlson, G. A. (2003). The bottom line. *Journal of Child and Adolescent Psychopharmacology, 13,* 115–118. http://dx.doi.org/10.1089/104454603322163826

Carlson, G. A. (2011). Will the child with mania please stand up? *The British Journal of Psychiatry, 198,* 171–172. http://dx.doi.org/10.1192/bjp.bp.110.084517

Carlson, G. A., & Kelly, K. L. (1998). Manic symptoms in psychiatrically hospitalized children—What do they mean? *Journal of Affective Disorders, 51,* 123–135.

Carlson, G. A., & Klein, D. N. (2014). How to understand divergent views on bipolar disorder in youth. *Annual Review of Clinical Psychology, 10,* 529–551. http://dx.doi.org/10.1146/annurev-clinpsy-032813-153702

*Carlson, G. A., Loney, J., Salisbury, H., & Volpe, R. J. (1998). Young referred boys with DICA-P manic symptoms vs. two comparison groups. *Journal of Affective Disorders, 51,* 113–121. http://dx.doi.org/10.1016/S0165-0327(98)00210-9

Carlson, G. A., & Youngstrom, E. A. (2003). Clinical implications of pervasive manic symptoms in children. *Biological Psychiatry, 53,* 1050–1058. http://dx.doi.org/10.1016/S0006-3223(03)00068-4

Carlson, G. A., & Youngstrom, E. A. (2011). Two opinions about one child—What's the clinician to do? *Journal of Child and Adolescent Psychopharmacology, 21,* 385–387. http://dx.doi.org/10.1089/cap.2011.2159

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18. http://dx.doi.org/10.1037/0022-006X.66.1.7

Cheung, F. M., & Leung, K. (1998). Indigenous personality measures: Chinese examples. *Journal of Cross-Cultural Psychology, 29,* 233–248. http://dx.doi.org/10.1177/0022022198291012

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conners, C. K. (1999). Conners Rating Scales-Revised. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 467–495). Mahwah, NJ: Erlbaum.

Cuthbert, B. N., & Insel, T. R. (2010). Toward new approaches to psychotic disorders: The NIMH Research Domain Criteria project. *Schizophrenia Bulletin, 36,* 1061–1062. http://dx.doi.org/10.1093/schbul/sbq108

Danielson, C. K., Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Discriminative validity of the general behavior inventory using youth report. *Journal of Abnormal Child Psychology, 31,* 29–39. http://dx.doi.org/10.1023/A:1021717231272

Delbello, M. P., Lopez-Larson, M. P., Soutullo, C. A., & Strakowski, S. M. (2001). Effects of race on psychiatric diagnosis of hospitalized adolescents: A retrospective chart review. *Journal of Child and Adolescent Psychopharmacology, 11,* 95–103. http://dx.doi.org/10.1089/104454601750143528

Dell'Osso, L., Pini, S., Cassano, G. B., Mastrocinque, C., Seckinger, R. A., Saettoni, M., . . . Amador, X. F. (2002). Insight into illness in patients with mania, mixed mania, bipolar depression and major depression with psychotic features. *Bipolar Disorders, 4,* 315–322. http://dx.doi.org/10.1034/j.1399-5618.2002.01192.x

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131,* 483–509. http://dx.doi.org/10.1037/0033-2909.131.4.483

Depue, R. A., Slater, J. F., Wolfstetter-Kausch, H., Klein, D., Goplerud, E., & Farr, D. (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. http://dx.doi.org/10.1037/0021-843X.90.5.381

de Sousa Gurgel, W., Rebouças, D. B., Negreiros de Matos, K. J., Carneiro, A. H., & Gomes de Matos e Souza, F., & the Grupo de Estudos em Transtornos Afetivos Affective Disorders Study Group. (2012). Brazilian Portuguese validation of Mood Disorder Questionnaire. *Comprehensive Psychiatry, 53,* 308–312. http://dx.doi.org/10.1016/j.comppsych.2011.04.059

*Dienes, K. A., Chang, K. D., Blasey, C. M., Adleman, N. E., & Steiner, H. (2002). Characterization of children of bipolar parents by parent report CBCL. *Journal of Psychiatric Research, 36,* 337–345. http://dx.doi.org/10.1016/S0022-3956(02)00019-5

*Diler, R. S., Birmaher, B., Axelson, D., Goldstein, B., Gill, M., Strober, M., . . . Keller, M. B. (2009). The Child Behavior Checklist (CBCL) and the CBCL-bipolar phenotype are not useful in diagnosing pediatric bipolar

disorder. *Journal of Child and Adolescent Psychopharmacology, 19,* 23–30. http://dx.doi.org/10.1089/cap.2008.067

*Doerfler, L. A., Connor, D. F., & Toscano, P. F. (2011). The CBCL bipolar profile and attention, mood, and behavior dysregulation. *Journal of Child and Family Studies, 20,* 545–553. http://dx.doi.org/10.1007/s10826-010-9426-z

Doerfler, L. A., Connor, D. F., & Toscano, P. F., Jr. (2011). Aggression, ADHD symptoms, and dysphoria in children and adolescents diagnosed with bipolar disorder and ADHD. *Journal of Affective Disorders, 131,* 312–319. http://dx.doi.org/10.1016/j.jad.2010.11.029

Drotar, D., Stein, R. E. K., & Perrin, E. C. (1995). Methodological issues in using the Child Behavior Checklist and its related instruments in clinical child psychology research. *Journal of Clinical Child Psychology, 24,* 184–192. http://dx.doi.org/10.1207/s15374424jccp2402_6

Dubicka, B., Carlson, G. A., Vail, A., & Harrington, R. (2008). Prepubertal mania: Diagnostic differences between US and UK clinicians. *European Child & Adolescent Psychiatry, 17,* 153–161. http://dx.doi.org/10.1007/s00787-007-0649-5

Du Rocher Schudlich, T. D., Youngstrom, E. A., Calabrese, J. R., & Findling, R. L. (2008). The role of family functioning in bipolar disorder in families. *Journal of Abnormal Child Psychology, 36,* 849–863. http://dx.doi.org/10.1007/s10802-008-9217-9

Ehnvall, A., Mitchell, P. B., Hadzi-Pavlovic, D., Loo, C., Breakspear, M., Wright, A., . . . Corry, J. (2011). Pain and rejection sensitivity in bipolar depression. *Bipolar Disorders, 13,* 59–66. http://dx.doi.org/10.1111/j.1399-5618.2011.00892.x

Emery, R. (1992). Family conflicts and their developmental implications: A conceptual analysis of meanings for the structure of relationships. In W. Hartup & C. Shantz (Eds.), *Family conflicts* (pp. 270–298). New York, NY: Cambridge.

Eyberg, S. M., Nelson, M. M., & Boggs, S. R. (2008). Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. *Journal of Clinical Child and Adolescent Psychology, 37,* 215–237. http://dx.doi.org/10.1080/15374410701820117

*Faraone, S. V., Althoff, R. R., Hudziak, J. J., Monuteaux, M., & Biederman, J. (2005). The CBCL predicts *DSM* bipolar disorder in children: A receiver operating characteristic curve analysis. *Bipolar Disorders, 7,* 518–524. http://dx.doi.org/10.1111/j.1399-5618.2005.00271.x

Findling, R. L., Frazier, T. W., Youngstrom, E. A., McNamara, N. K., Stansbrey, R. J., Gracious, B. L., . . . Calabrese, J. R. (2007). Double-blind, placebo-controlled trial of divalproex monotherapy in the treatment of symptomatic youth at high risk for developing bipolar disorder. *The Journal of Clinical Psychiatry, 68,* 781–788. http://dx.doi.org/10.4088/JCP.v68n0519

Findling, R. L., McNamara, N. K., Youngstrom, E. A., Stansbrey, R., Gracious, B. L., Reed, M. D., & Calabrese, J. R. (2005). Double-blind 18-month trial of lithium versus divalproex maintenance treatment in pediatric bipolar disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 44,* 409–417. http://dx.doi.org/10.1097/01.chi.0000155981.83865.ea

*Findling, R. L., Youngstrom, E. A., Fristad, M. A., Birmaher, B., Kowatch, R. A., Arnold, L. E., . . . Horwitz, S. M. (2010). Characteristics of children with elevated symptoms of mania: The Longitudinal Assessment of Manic Symptoms (LAMS) study. *Journal of Clinical Psychiatry, 71,* 1664–1672. http://dx.doi.org/10.4088/JCP.09m05859yel

*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. http://dx.doi.org/10.1111/j.1399-5618.2005.00260.x

Freeman, A. J., Youngstrom, E. A., Freeman, M. J., Youngstrom, J. K., & Findling, R. L. (2011). Is caregiver-adolescent disagreement due to differences in thresholds for reporting manic symptoms? *Journal of Child and Adolescent Psychopharmacology, 21,* 425–432. http://dx.doi.org/10.1089/cap.2011.0033

Fristad, M. A., & MacPherson, H. A. (2014). Evidence-based psychosocial treatments for child and adolescent bipolar spectrum disorders. *Journal of Clinical Child and Adolescent Psychology, 43,* 339–355. http://dx.doi.org/10.1080/15374416.2013.822309

Gadow, K. D., & Sprafkin, J. (1994). *Child symptom inventories manual*. Stony Brook, NY: Checkmate Plus.

Gadow, K. D., & Sprafkin, J. (1997). *Adolescent Symptom Inventory: Screening manual*. Stony Brook, NY: Checkmate Plus.

Galanter, C. A., Hundt, S. R., Goyal, P., Le, J., & Fisher, P. W. (2012). Variability among research diagnostic interview instruments in the application of *DSM–IV–TR* criteria for pediatric bipolar disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 51,* 605–621. http://dx.doi.org/10.1016/j.jaac.2012.03.010

Gamma, A., Angst, J., Azorin, J. M., Bowden, C. L., Perugi, G., Vieta, E., & Young, A. H. (2013). Transcultural validity of the Hypomania Checklist-32 (HCL-32) in patients with major depressive episodes. *Bipolar Disorders, 15,* 701–712. http://dx.doi.org/10.1111/bdi.12101

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10299-000

Geller, B., & DelBello, M. P. (Eds.). (2003). *Bipolar disorder in childhood and early adolescence*. New York, NY: Guilford Press.

Geller, B., & Luby, J. (1997). Child and adolescent bipolar disorder: A review of the past 10 years. *Journal of the American Academy of Child & Adolescent Psychiatry, 36,* 1168–1176. http://dx.doi.org/10.1097/00004583-199709000-00008

Geller, B., & Tillman, R. (2005). Prepubertal and early adolescent bipolar I disorder: Review of diagnostic validation by Robins and Guze criteria. *Journal of Clinical Psychiatry, 66*(Suppl. 7), 21–28.

*Geller, B., Warner, K., Williams, M., & Zimerman, B. (1998). Prepubertal and young adolescent bipolarity versus ADHD: Assessment and validity using the WASH-U-KSADS, CBCL and TRF. *Journal of Affective Disorders, 51,* 93–100. http://dx.doi.org/10.1016/S0165-0327(98)00176-1 AQ: 8

Geller, B., Williams, M., Zimerman, B., Frazier, J., Beringer, L., & Warner, K. L. (1998). Prepubertal and early adolescent bipolarity differentiate from ADHD by manic symptoms, grandiose delusions, ultra-rapid or ultradian cycling. *Journal of Affective Disorders, 51,* 81–91. http://dx.doi.org/10.1016/S0165-0327(98)00175-X

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. http://dx.doi.org/10.1097/00004583-200104000-00014

Gore, F. M., Bloem, P. J., Patton, G. C., Ferguson, J., Joseph, V., Coffey, C., . . . Mathers, C. D. (2011). Global burden of disease in young people aged 10–24 years: A systematic analysis. *Lancet, 377,* 2093–2102.

Gracious, B. L., Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2002). Discriminative validity of a parent version of the Young Mania Rating Scale. *Journal of the American Academy of Child & Adolescent Psychiatry, 41,* 1350–1359. http://dx.doi.org/10.1097/00004583-200211000-00017

Hartley, C. A., & Phelps, E. A. (2010). Changing fear: The neurocircuitry of emotion regulation. *Neuropsychopharmacology, 35,* 136–146. http://dx.doi.org/10.1038/npp.2009.121

Harvey, A. G. (2008). Sleep and circadian rhythms in bipolar disorder: Seeking synchrony, harmony, and regulation. *The American Journal of Psychiatry, 165,* 820–829. http://dx.doi.org/10.1176/appi.ajp.2008.08010098

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117,* 167–178. http://dx.doi.org/10.1037/0033-2909.117.1.167

Hauser, M., & Correll, C. U. (2013). The significance of at-risk or prodromal symptoms for bipolar I disorder in children and adolescents. *The Canadian Journal of Psychiatry, 58,* 22–31.

*Hazell, P. L., Lewin, T. J., & Carr, V. J. (1999). Confirmation that Child Behavior Checklist clinical scales discriminate juvenile mania from attention deficit hyperactivity disorder. *Journal of Paediatrics and Child Health, 35,* 199–203. http://dx.doi.org/10.1046/j.1440-1754.1999.t01-1-00347.x

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203–217. http://dx.doi.org/10.1037/1082-989X.6.3.203

Henin, A., Mick, E., Biederman, J., Fried, R., Wozniak, J., Faraone, S. V., . . . Doyle, A. E. (2007). Can bipolar disorder-specific neuropsychological im-

pairments in children be identified? *Journal of Consulting and Clinical Psychology, 75,* 210–220. http://dx.doi.org/10.1037/0022-006X.75.2.210

*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. http://dx.doi.org/10.1002/jclp.20464

Hirschfeld, R. M., Williams, J. B. W., Spitzer, R. L., Calabrese, J. R., Flynn, L., Keck, P. E. J., 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. http://dx.doi.org/10.1176/appi.ajp.157.11.1873

Hummel, T. J. (1999). The usefulness of tests in clinical decisions. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 59–112). Boston, MA: Allyn & Bacon.

Jenkins, M. M., Youngstrom, E. A., Washburn, J. J., & Youngstrom, J. K. (2011). Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice, 42,* 121–129. http://dx.doi.org/10.1037/a0022506

Jenkins, M. M., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2012). Generalizability of evidence-based assessment recommendations for pediatric bipolar disorder. *Psychological Assessment, 24,* 269–281. http://dx.doi.org/10.1037/a0025775

Jensen-Doss, A., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2014). Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. *Journal of Consulting and Clinical Psychology, 82,* 1151–1162. http://dx.doi.org/10.1037/a0036657

Johnson, S. L., Miller, C. J., & Eisner, L. (2008). Bipolar Disorder. In J. Hunsley & E. J. Mash (Eds.), *A guide to assessments that work* (pp. 121–137). New York, NY: Oxford University Press.

Joseph, M. F., Frazier, T. W., Youngstrom, E. A., & Soares, J. C. (2008). A quantitative and qualitative review of neurocognitive performance in pediatric bipolar disorder. *Journal of Child and Adolescent Psychopharmacology, 18,* 595–605. http://dx.doi.org/10.1089/cap.2008.064

Kahana, S. Y., Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Employing parent, teacher, and youth self-report checklists in identifying pediatric bipolar spectrum disorders: An examination of diagnostic accuracy and clinical utility. *Journal of Child and Adolescent Psychopharmacology, 13,* 471–488. http://dx.doi.org/10.1089/104454603322724869

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. http://dx.doi.org/10.1097/00004583-199707000-00021

Keck, P. E. J., Jr., Welge, J. A., Strakowski, S. M., Arnold, L. M., & McElroy, S. L. (2000). Placebo effect in randomized, controlled maintenance studies of patients with bipolar disorder. *Biological Psychiatry, 47,* 756–761. http://dx.doi.org/10.1016/S0006-3223(99)00309-1

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2,* 61–76. http://dx.doi.org/10.1002/jrsm.35

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. http://dx.doi.org/10.1111/j.1399-5618.2005.00261.x

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

*Lee, H. J., Joo, Y., Youngstrom, E. A., Yum, S. Y., Findling, R. L., & Kim, H. W. (2014). Diagnostic validity and reliability of a Korean version of the Parent and Adolescent General Behavior Inventories. *Comprehensive Psychiatry, 55,* 1730–1737. http://dx.doi.org/10.1016/j.comppsych.2014.05.008

Leibenluft, E. (2011). Severe mood dysregulation, irritability, and the diagnostic boundaries of bipolar disorder in youths. *The American Journal of Psychiatry, 168,* 129–142. http://dx.doi.org/10.1176/appi.ajp.2010.10050766

AQ: 9

Leibenluft, E., Charney, D. S., Towbin, K. E., Bhangoo, R. K., & Pine, D. S. (2003). Defining clinical phenotypes of juvenile mania. *The American Journal of Psychiatry, 160,* 430–437. http://dx.doi.org/10.1176/appi.ajp.160.3.430

*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.

Lewinsohn, P. M., Klein, D. N., & Seeley, J. R. (2000). Bipolar disorder during adolescence and young adulthood in a community sample. *Bipolar Disorders, 2,* 281–293. http://dx.doi.org/10.1034/j.1399-5618.2000.20309.x

Lewinsohn, P., Seeley, J. R., & Klein, D. N. (2003). Bipolar disorder in adolescents: Epidemiology and suicidal behavior. In B. Geller & M. P. DelBello (Eds.), *Bipolar disorder in childhood and early adolescence* (pp. 7–24). New York, NY: Guilford Press.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *British Medical Journal, 339,* b2700. http://dx.doi.org/10.1136/bmj.b2700

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage.

Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology, 19,* 136–143. http://dx.doi.org/10.1207/s15374424jccp1902_5

Mackin, P., Targum, S. D., Kalali, A., Rom, D., & Young, A. H. (2006). Culture and assessment of manic symptoms. *The British Journal of Psychiatry, 189,* 379–380. http://dx.doi.org/10.1192/bjp.bp.105.013920

*Marchand, W. R., Clark, S. C., Wirth, L., & Simon, C. (2005). Validity of the parent young mania rating scale in a community mental health setting. *Psychiatry, 2,* 31–35.

Mbekou, V., Gignac, M., MacNeil, S., Mackay, P., & Renaud, J. (2014). The CBCL dysregulated profile: An indicator of pediatric bipolar disorder or of psychopathology severity? *Journal of Affective Disorders, 155,* 299–302. http://dx.doi.org/10.1016/j.jad.2013.10.033

McClellan, J., Kowatch, R., Findling, R. L., & the Work Group on Quality Issues. (2007). Practice parameter for the assessment and treatment of children and adolescents with bipolar disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 46,* 107–125. http://dx.doi.org/10.1097/01.chi.0000242240.69678.c4

Meeuwesen, L., van den Brink-Muinen, A., & Hofstede, G. (2009). Can dimensions of national culture predict cross-national differences in medical communication? *Patient Education and Counseling, 75,* 58–66. http://dx.doi.org/10.1016/j.pec.2008.09.015

Merikangas, K. R., Akiskal, H. S., Angst, J., Greenberg, P. E., Hirschfeld, R. M. A., Petukhova, M., & Kessler, R. C. (2007). Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of General Psychiatry, 64,* 543–552. http://dx.doi.org/10.1001/archpsyc.64.5.543

Merikangas, K. R., He, J. P., Burstein, M., Swendsen, J., Avenevoli, S., Case, B., . . . Olfson, M. (2011). Service utilization for lifetime mental disorders in U.S. adolescents: Results of the National Comorbidity Survey-Adolescent Suppl. (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry, 50,* 32–45. http://dx.doi.org/10.1016/j.jaac.2010.10.006

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. http://dx.doi.org/10.1111/j.1468-2850.2009.01152.x

Meyer, G. J. (2003). Guidelines for reporting information in studies of diagnostic test accuracy: The STARD initiative. *Journal of Personality Assessment, 81,* 191–193. http://dx.doi.org/10.1207/S15327752JPA8103_01

*Meyer, S. E., Carlson, G. A., Youngstrom, E., Ronsaville, D. S., Martinez, P. E., Gold, P. W., . . . Radke-Yarrow, M. (2009). Long-term outcomes of youth who manifested the CBCL-Pediatric Bipolar Disorder phenotype during childhood and/or adolescence. *Journal of Affective Disorders, 113,* 227–235. http://dx.doi.org/10.1016/j.jad.2008.05.024

Meyer, T. D., Hammelstein, P., Nilsson, L. G., Skeppar, P., Adolfsson, R., & Angst, J. (2007). The Hypomania Checklist (HCL-32): Its factorial structure and association to indices of impairment in German and Swedish nonclinical

samples. *Comprehensive Psychiatry, 48,* 79–87. http://dx.doi.org/10.1016/j.comppsych.2006.07.001

Mick, E., Biederman, J., Pandina, G., & Faraone, S. V. (2003). A preliminary meta-analysis of the child behavior checklist in pediatric bipolar disorder. *Biological Psychiatry, 53,* 1021–1027. http://dx.doi.org/10.1016/S0006-3223(03)00234-8

*Miguez, M., Weber, B., Debbane, M., Balanzin, D., Gex-Fabry, M., Raiola, F., . . . Aubry, J. M. (2012). Screening for bipolar disorder in adolescents with the Mood Disorder Questionnaire-Adolescent version (MDQ-A) and the Child Bipolar Questionnaire (CBQ). [Advance online publication]. *Early Intervention in Psychiatry*.

Miklowitz, D. J. (2002). *The bipolar disorder survival guide: What you and your family need to know*. New York, NY: Guilford Press.

Miller, C. J., Johnson, S. L., Kwapil, T. R., & Carver, C. S. (2011). Three studies on self-report scales to detect bipolar disorder. *Journal of Affective Disorders, 128,* 199–210. http://dx.doi.org/10.1016/j.jad.2010.07.012

Minkov, M., & Hofstede, G. (2011). The evolution of Hofstede's doctrine. *Cross Cultural Management: An International Journal, 18,* 10–20. http://dx.doi.org/10.1108/13527601111104269

Morrison, J. (2007). *Diagnosis made easier: Principles and techniques for mental health clinicians*. New York, NY: Guilford Press.

Murray, G., Nicholas, C. L., Kleiman, J., Dwyer, R., Carrington, M. J., Allen, N. B., & Trinder, J. (2009). Nature's clocks and human mood: The circadian system modulates reward motivation. *Emotion, 9,* 705–716. http://dx.doi.org/10.1037/a0017080

Nottelmann, E. (2001). National Institute of Mental Health research roundtable on prepubertal bipolar disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 40,* 871–878. http://dx.doi.org/10.1097/00004583-200108000-00007

Orvaschel, H. (1995). *Schizophrenia and affective disorders schedule for children-Epidemiological version (KSADS-E)*. Ft. Lauderdale, FL: Nova Southeastern University.

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157–159. http://dx.doi.org/10.2307/1164923

*Papachristou, E., Ormel, J., Oldehinkel, A. J., Kyriakopoulos, M., Reinares, M., Reichenberg, A., & Frangou, S. (2013). Child Behavior Checklist-Mania Scale (CBCL-MS): Development and evaluation of a population-based screening scale for bipolar disorder. *PLoS ONE, 8,* e69459.

Papolos, D. F. (2003). Bipolar disorder and comorbid disorders: The case for a dimensional nosology. In B. Geller & M. P. DelBello (Eds.), *Bipolar disorder in childhood and early adolescence* (pp. 76–106). New York, NY: Guilford Press.

*Papolos, D., Hennen, J., Cockerham, M. S., Thode, H. C. J., Jr., & Youngstrom, E. A. (2006). The child bipolar questionnaire: A dimensional approach to screening for pediatric bipolar disorder. *Journal of Affective Disorders, 95,* 149–158. http://dx.doi.org/10.1016/j.jad.2006.03.026

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. http://dx.doi.org/10.1097/01.chi.0000205700.40700.50

Pendergast, L. L., Youngstrom, E. A., Brown, C., Jensen, D., Abramson, L. Y., & Alloy, L. B. (2015). Structural invariance of General Behavior Inventory (GBI) scores in Black and White young adults. *Psychological Assessment, 27,* 21–30. http://dx.doi.org/10.1037/pas0000020

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63,* 539–569. http://dx.doi.org/10.1146/annurev-psych-120710-100452

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reichart, C. G., van der Ende, J., Wals, M., Hillegers, M. H., Nolen, W. A., Ormel, J., & Verhulst, F. C. (2005). The use of the GBI as predictor of bipolar disorder in a population of adolescent offspring of parents with a bipolar disorder. *Journal of Affective Disorders, 89,* 147–155. http://dx.doi.org/10.1016/j.jad.2005.09.007

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. http://dx.doi.org/10.1002/mpr.289

Richters, J. E. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin, 112,* 485–499. http://dx.doi.org/10.1037/0033-2909.112.3.485

Robertson, H. A., Lam, R. W., Stewart, J. N., Yatham, L. N., Tam, E. M., & Zis, A. P. (1996). Atypical depressive symptoms and clusters in unipolar and bipolar depression. *Acta Psychiatrica Scandinavica, 94,* 421–427. http://dx.doi.org/10.1111/j.1600-0447.1996.tb09884.x

Rocha-Rego, V., Jogia, J., Marquand, A. F., Mourao-Miranda, J., Simmons, A., & Frangou, S. (2014). Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: A pattern classification approach. *Psychological Medicine, 44,* 519–532. http://dx.doi.org/10.1017/S0033291713001013

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution; International Journal of Organic Evolution, 59,* 464–468. http://dx.doi.org/10.1111/j.0014-3820.2005.tb01004.x

*Rucklidge, J. J. (2008). Retrospective parent report of psychiatric histories: Do checklists reveal specific prodromal indicators for postpubertal-onset pediatric bipolar disorder? *Bipolar Disorders, 10,* 56–66. http://dx.doi.org/10.1111/j.1399-5618.2008.00533.x

Sattler, J. M. (2002). *Assessment of children: Behavioral and clinical applications* (4th ed.). La Mesa, CA: Author.

Scheffer, R. E., Kowatch, R. A., Carmody, T., & Rush, A. J. (2005). Randomized, placebo-controlled trial of mixed amphetamine salts for symptoms of comorbid ADHD in pediatric bipolar disorder after mood stabilization with divalproex sodium. *The American Journal of Psychiatry, 162,* 58–64. http://dx.doi.org/10.1176/appi.ajp.162.1.58

*Serrano, E., Ezpeleta, L., Alda, J. A., Matalí, J. L., & San, L. (2011). Psychometric properties of the Young Mania Rating Scale for the identification of mania symptoms in Spanish children and adolescents with attention deficit/hyperactivity disorder. *Psychopathology, 44,* 125–132. http://dx.doi.org/10.1159/000320893

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

Stedman, J. M., Hatch, J. P., & Schoenfeld, L. S. (2001). The current status of psychological assessment training in graduate and professional schools. *Journal of Personality Assessment, 77,* 398–407. http://dx.doi.org/10.1207/S15327752JPA7703_02

Stewart, A. J., Theodore-Oklota, C., Hadley, W., Brown, L. K., Donenberg, G., & DiClemente, R., & the Project STYLE Study Group. (2012). Mania symptoms and HIV-risk behavior among adolescents in mental health treatment. *Journal of Clinical Child and Adolescent Psychology, 41,* 803–810. http://dx.doi.org/10.1080/15374416.2012.675569

Strakowski, S. M., Adler, C. M., Almeida, J., Altshuler, L. L., Blumberg, H. P., Chang, K. D., . . . Townsend, J. D. (2012). The functional neuroanatomy of bipolar disorder: A consensus model. *Bipolar Disorders, 14,* 313–325.

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-based medicine: How to practice and teach EBM* (4th ed.). New York, NY: Churchill Livingstone.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26. http://dx.doi.org/10.1111/1529-1006.001

*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. http://dx.doi.org/10.1176/appi.ajp.162.6.1214

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9,* 151–176. http://dx.doi.org/10.1146/annurev-clinpsy-050212-185510

*Uchida, M., Faraone, S. V., Martelon, M., Kenworthy, T., Woodworth, K. Y., Spencer, T. J., . . . Biederman, J. (2014). Further evidence that severe scores in the aggression/anxiety-depression/attention subscales of child behavior checklist (severe dysregulation profile) can screen for bipolar disorder symptomatology: A conditional probability analysis. *Journal of Affective Disorders, 165,* 81–86. http://dx.doi.org/10.1016/j.jad.2014.04.021

Van Meter, A. R., Moreira, A. L., & Youngstrom, E. A. (2011). Meta-analysis of epidemiologic studies of pediatric bipolar disorder. *Journal of Clinical Psychiatry, 72,* 1250–1256. http://dx.doi.org/10.4088/JCP.10m06290

Van Meter, A., Youngstrom, E. A., Demeter, C., & Findling, R. L. (2013). Examining the validity of cyclothymic disorder in a youth sample: Replication and extension. *Journal of Abnormal Child Psychology, 41,* 367–378. http://dx.doi.org/10.1007/s10802-012-9680-1

Van Meter, A. R., Youngstrom, E. A., & Findling, R. L. (2012). Cyclothymic disorder: A critical review. *Clinical Psychology Review, 32,* 229–243. http://dx.doi.org/10.1016/j.cpr.2012.02.001

Van Meter, A., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2011). Examining the validity of cyclothymic disorder in a youth sample. *Journal of Affective Disorders, 132,* 55–63. http://dx.doi.org/10.1016/j.jad.2011.02.004

Van Meter, A., Youngstrom, E., Youngstrom, J. K., Ollendick, T., Demeter, C., & Findling, R. L. (2014). Clinical decision making about child and adolescent anxiety disorders using the Achenbach system of empirically based assessment. *Journal of Clinical Child and Adolescent Psychology, 43,* 552–565. http://dx.doi.org/10.1080/15374416.2014.883930

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics, 56,* 1134–1138. http://dx.doi.org/10.1111/j.0006-341X.2000.01134.x

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 60,* 29–60. http://dx.doi.org/10.1348/000711005X64042

Viechtbauer, W. (2010a). Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software, 36,* 1–48. http://dx.doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2010b). *Metafor: meta-analysis package for R. R package version,* 2010, 1–0.

Vieta, E., Reinares, M., & Rosa, A. R. (2011). Staging bipolar disorder. *Neurotoxicity Research, 19,* 279–285. http://dx.doi.org/10.1007/s12640-010-9197-8

*Wagner, K. D., Hirschfeld, R. M., 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. http://dx.doi.org/10.4088/JCP.v67n0518

Wakschlag, L. S., Choi, S. W., Carter, A. S., Hullsiek, H., Burns, J., McCarthy, K., . . . Briggs-Gowan, M. J. (2012). Defining the developmental parameters of temper loss in early childhood: Implications for developmental psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 53,* 1099–1108. http://dx.doi.org/10.1111/j.1469-7610.2012.02595.x

Walshaw, P. D., Alloy, L. B., & Sabb, F. W. (2010). Executive function in pediatric bipolar disorder and attention-deficit hyperactivity disorder: In search of distinct phenotypic profiles. *Neuropsychology Review, 20,* 103–120. http://dx.doi.org/10.1007/s11065-009-9126-x

Waugh, M. J., Meyer, T. D., Youngstrom, E. A., & Scott, J. (2014). A review of self-rating instruments to identify young people at risk of bipolar spectrum disorders. *Journal of Affective Disorders, 160,* 113–121. http://dx.doi.org/10.1016/j.jad.2013.12.019

West, A. E., Celio, C. I., Henry, D. B., & Pavuluri, M. N. (2011). Child Mania Rating Scale-Parent Version: A valid measure of symptom change due to pharmacotherapy. *Journal of Affective Disorders, 128,* 112–119. http://dx.doi.org/10.1016/j.jad.2010.06.013

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., . . . Bossuyt, P. M., & the QUADAS-2 Group. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155,* 529–536. http://dx.doi.org/10.7326/0003-4819-155-8-201110180-00009

Wilens, T. E., Biederman, J., Forkner, P., Ditterline, J., Morris, M., Moore, H., . . . Wozniak, J. (2003). Patterns of comorbidity and dysfunction in clinically referred preschool and school-age children with bipolar disorder. *Journal of Child and Adolescent Psychopharmacology, 13,* 495–505. http://dx.doi.org/10.1089/104454603322724887

Wozniak, J., Biederman, J., Kiely, K., Ablon, J. S., Faraone, S. V., Mundy, E., & Mennin, D. (1995). Mania-like symptoms suggestive of childhood-onset bipolar disorder in clinically referred children. *Journal of the American Academy of Child & Adolescent Psychiatry, 34,* 867–876. http://dx.doi.org/10.1097/00004583-199507000-00010

Yatham, L. N., Kennedy, S. H., O'Donovan, C., Parikh, S., MacQueen, G., McIntyre, R., . . . Gorman, C. P. (2005). Canadian Network for Mood and Anxiety Treatments (CANMAT) guidelines for the management of patients with bipolar disorder: Consensus and controversies. *Bipolar Disorders, 7* (*Suppl. 3), 5–69. http://dx.doi.org/10.1111/j.1399-5618.2005.00219.x*

Yeh, M., & Weisz, J. R. (2001). Why are we here at the clinic? Parent-child (dis)agreement on referral problems at outpatient treatment entry. *Journal of Consulting and Clinical Psychology, 69,* 1018–1025. http://dx.doi.org/10.1037/0022-006X.69.6.1018

You, D. S., Youngstrom, E. A., Feeny, N. C., Youngstrom, J. K., & Findling, R. L. (2015). Comparing the diagnostic accuracy of five instruments for detecting posttraumatic stress disorder in youth. [Advance online publication]. *Journal of Clinical Child and Adolescent Psychology,* 1–12. http://dx.doi.org/10.1080/15374416.2015.1030754

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

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

Youngstrom, E. A. (2009). Definitional issues in bipolar disorder across the life cycle. *Clinical Psychology: Science and Practice, 16,* 140–160. http://dx.doi.org/10.1111/j.1468-2850.2009.01154.x

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology, 39,* 204–221. http://dx.doi.org/10.1093/jpepsy/jst062

Youngstrom, E. A., Arnold, L. E., & Frazier, T. W. (2010). Bipolar and ADHD comorbidity: Both artifact and outgrowth of shared mechanisms. *Clinical Psychology: Science and Practice, 17,* 350–359. http://dx.doi.org/10.1111/j.1468-2850.2010.01226.x

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

Youngstrom, E. A., Choukas-Bradley, S., Calhoun, C. D., & Jensen-Doss, A. (2015). Clinical guide to the Evidence-Based Assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice, 22,* 20–35. http://dx.doi.org/10.1016/j.cbpra.2013.12.005

Youngstrom, E. A., & De Los Reyes, A. (2015). Commentary: Moving toward cost-effectiveness in using psychophysiological measures in clinical assessment: Validity, decision making, and adding value. *Journal of Clinical Child & Adolescent Psychology, 44,* 352–361. http://dx.doi.org/10.1080/15374416.2014.913252

Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Who are the comorbid adolescents? Agreement between psychiatric diagnosis, youth, parent, and teacher report. *Journal of Abnormal Child Psychology, 31,* 231–245. http://dx.doi.org/10.1023/A:1023244512119

Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2004). Effects of adolescent manic symptoms on agreement between youth, parent, and teacher ratings of behavior problems. *Journal of Affective Disorders, 82*(Suppl. 1), S5–S16. http://dx.doi.org/10.1016/j.jad.2004.05.016

Youngstrom, E. A., Findling, R. L., Calabrese, J. R., Gracious, B. L., Demeter, C., Bedoya, D. D., & Price, M. (2004). Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *Journal of the American Academy of Child & Adolescent Psychiatry, 43,* 847–858. http://dx.doi.org/10.1097/01.chi.0000125091.35109.1e

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. http://dx.doi.org/10.1037/1040-3590.13.2.267

Youngstrom, E. A., Jenkins, M. M., Jensen-Doss, A., & Youngstrom, J. K. (2012). Evidence-based assessment strategies for pediatric bipolar disorder. *The Israel Journal of Psychiatry and Related Sciences, 49,* 15–27.

Youngstrom, E. A., Joseph, M. F., & Greene, J. (2008). Comparing the psychometric properties of multiple teacher report instruments as predictors of bipolar disorder in children and adolescents. *Journal of Clinical Psychology, 64,* 382–401. http://dx.doi.org/10.1002/jclp.20462

*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. http://dx.doi.org/10.1111/j.1399-5618.2005.00269.x

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

Youngstrom, E., Meyers, O., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006b). Diagnostic and measurement issues in the assessment of pediatric bipolar disorder: Implications for understanding mood disorder across the life cycle. *Development and Psychopathology, 18,* 989–1021. http://dx.doi.org/10.1017/S0954579406060494

Youngstrom, E. A., Youngstrom, J. K., Freeman, A. J., De Los Reyes, A., Feeny, N. C., & Findling, R. L. (2011). Informants are not all equal: Predictors and correlates of clinician judgments about caregiver and youth credibility. *Journal of Child and Adolescent Psychopharmacology, 21,* 407–415. http://dx.doi.org/10.1089/cap.2011.0032

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

Youngstrom, E. A. (2015). *Raw data for meta-analysis of discriminative validity of caregiver, youth, and teacher report for pediatric bipolar disorder—All English publications through End of 2014. ICPSR36245-v1.* Ann Arbor, MI: Inter-university Consortium for Political and Social Research. http://dx.doi.org/10.3886/ICPSR36245.v1

Zaratiegui, R. M., Vázquez, G. H., Lorenzo, L. S., Marinelli, M., Aguayo, S., Strejilevich, S. A., . . . Ghaemi, N. (2011). Sensitivity and specificity of the mood disorder questionnaire and the bipolar spectrum diagnostic scale in Argentinean patients with mood disorders. *Journal of Affective Disorders, 132,* 445–449. http://dx.doi.org/10.1016/j.jad.2011.03.014

Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York, NY: Wiley. http://dx.doi.org/10.1002/9780470317082

Received May 13, 2015
Revision received July 14, 2015
Accepted July 17, 2015 ■