Published in final edited form as:

J Clin Child Adolesc Psychol. 2018 ; 47(5): 832–846. doi:10.1080/15374416.2017.1280807.

Diagnostic Efficiency of the Child and Adolescent Symptom Inventory (CASI-4R) Depression Subscales for identifying Youth Mood Disorders

Stephanie Salcedo¹, Yen-Ling Chen², Eric A. Youngstrom¹, Mary A. Fristad³, Kenneth D. Gadow⁴, Sarah M. Horwitz⁵, Thomas W. Frazier⁶, L. Eugene Arnold³, Mary L. Phillips⁷, Boris Birmaher⁷, Robert A. Kowatch⁸, and Robert L. Findling⁹

¹University of North Carolina, Chapel Hill

²University of Nevada, Las Vegas

³Ohio State University

⁴Stony Brook University Department of Psychiatry

⁵New York University

⁶Cleveland Clinic

⁷University of Pittsburgh

⁸Cincinnati Children's Hospital Medical Center

⁹Johns Hopkins University/Kennedy Krieger Institute

Abstract

This study examined the diagnostic and clinical utility of the Child and Adolescent Symptom Inventory-4R (CASI-4R) depressive and dysthymia subscale for detecting mood disorders in youth (ages 6-12; $M = 9.37$) visiting outpatient mental health clinics. Secondary analyses ($N = 700$) utilized baseline data from the Longitudinal Assessment of Manic Symptoms (LAMS) study. Semi-structured interviews with youth participants and their caregivers determined psychiatric diagnoses. Caregivers and teachers completed the CASI-4R. CASI-4R depressive symptom severity and symptom count scores each predicted mood disorder diagnoses. Both caregiver versions of the CASI-4R subscale significantly identified youth mood disorders (areas under the curve [AUCs] = .78 - .79, $ps < .001$). The symptom severity version showed a small but significant advantage. Teacher symptom severity report did not significantly predict mood disorder diagnosis ($AUC = .56$, $p > .05$), whereas the teacher symptom count report corresponded to a small effect size ($AUC = .61$, $p < .05$). The CASI-4R depression scale showed strong incremental validity even controlling for the other CASI-4R scales. Caregiver subscale cutoff scores were calculated to assist in ruling in (diagnostic likelihood ratio [DLR] = 3.73) or ruling out (DLR = 0.18) presence of a mood disorder. The CASI-4R depressive subscale caregiver report can help identify youth mood disorders and using DLRs may help improve diagnostic accuracy.

Depression affects a significant number of children and adolescents: epidemiological studies report three percent lifetime depressive disorder prevalence in preadolescence and up to 25% experiencing at least one episode by the end of adolescence (Kessler, Avenevoli, & Ries

Merikangas, 2001; Lewinsohn & Essau, 2002). Twenty to 50 percent of youth self-report depressive symptoms at clinically significant levels (Kessler et al., 2001). Although there are no differences in prevalence by sex during childhood, by early adolescence, girls' rates increase dramatically, to up to two times higher by age 15 (Hankin et al., 1998). This sex difference remains throughout adulthood (Kessler et al., 2012). Individuals with depressive symptoms can experience a wide range of severe impairments, such as deficient academic performance, relationship conflicts, and negative self-concepts (Garber & Horowitz, 2002; Hammen & Rudolph, 2003; Lewinsohn & Essau, 2002).

Extensive evidence shows high rates of comorbidity of depression with other psychiatric disorders in youth. In particular, data show significant comorbidity between depression and anxiety disorders, conduct disorder, oppositional defiant disorder (ODD), attention deficit/hyperactivity disorder (ADHD) and substance use disorders (Angold, Costello, & Erkanli, 1999; Lewinsohn, Hops, Roberts, Seeley, & Andrews, 1993). Furthermore, depression is the leading risk factor in youth suicide (Birmaher, Arbelaez, & Brent, 2002). Youth with depression may still face impairment even following remission or successful treatment (Garber & Horowitz, 2002; Rao et al., 1995), and depression earlier in life predicts impairment into adulthood (Birmaher et al., 2002; Dunn & Goodyer, 2006; Lewinsohn, Rohde, Seeley, Klein, & Gotlib, 2003). Although most children and adolescents with a major depressive episode recover naturally (Essau, 2007), relapse rates are high, with the majority of youth experiencing another episode within several years later (Birmaher et al., 2002; Kovacs, 1996; Pettit, Hartley, Lewinsohn, Seeley, & Klein, 2013).

Given the multitude of consequences of suffering from depression, it is crucial to identify depressive symptoms in children as early as possible to begin treatment. However, both under diagnosis and under treatment persist, and there is often a lag between youth mood symptom onset, diagnosis and initiation of treatment (Kessler, Merikangas, & Wang, 2007).

Incorporating Multiple Informants

When assessing for psychopathology in youth, collateral reports from various informants (e.g., parents, teachers) and settings (e.g., home, classroom) help provide an accurate picture of the youth's presenting symptoms (De Los Reyes et al., 2015). The added value of cross-informant reports varies depending on the child's age (De Los Reyes & Kazdin, 2005); children are less reliable reporters of their psychological symptoms than adolescents, and adolescents in turn spend less time with any one particular teacher (Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985). Parent report may be more useful for younger children because parents are more involved in the everyday lives of their younger children than of their adolescents, so they may be able to provide more background knowledge about their behavior and activities (De Los Reyes et al., 2011; Richters & Pellegrini, 1989). Adolescents often have issues reporting on their symptom timeline and course, so parent and teacher report may still be useful (Edelbrock et al., 1985; Jensen et al., 1999). With elementary school-aged children, parent and teacher report may be essential for early identification of mood disorders, as reading ability and meta-cognitive development may preclude using self-report rating scales. Many widely used rating scales do not include youth younger than 11

years old in their standardization samples, and even then, only when items have simple reading levels (e.g., Achenbach & Rescorla, 2001).

Calculating Clinically Useful Effect Sizes for Diagnostic Efficiency

Brief assessments can help to determine whether someone is exhibiting symptoms consistent with a particular set of disorders. Intuitively, one would expect that youth with depression would score higher on a depression screening tool than youth without depression. However, what is less clear is how a score on a particular assessment tool influences an individual's relative risk for a disorder. Signal detection theory (McFall & Treat, 1999; Swets, Dawes, & Monahan, 2000) and Bayesian methods (Bayes & Price, 1763; Kruschke, 2011) provide the theoretical and statistical framework to translate research data into information that clinicians can utilize to answer this question, specifically via use of Receiver Operator Characteristics (ROC).

ROC evaluates test scores as predictors of a clinically meaningful dichotomous variable (e.g., major depressive disorder diagnosis). It provides the foundation to apply group data to individual cases to examine diagnostic probability, which could then inform clinical decision-making. ROC utilizes a test's *sensitivity* and *specificity* to evaluate diagnostic accuracy. Sensitivity is the proportion of individuals with a disorder who also received a positive test result for that disorder; specificity is the proportion of individuals without the disorder who received a negative test result. The *false positive rate* is equal to $1 - \text{specificity}$ (Youngstrom, 2014). ROC plots the sensitivity by the false positive rate of a particular test; a test with zero discriminating power would have a ROC plot around a 45-degree diagonal line, and the better the test is able to discriminate between individuals with the disorder from those without it, the further its ROC curve will move toward the upper-left corner of the graph. ROC curves can be quantified by calculating the area under the curve (AUC), which is an effect size. AUCs of two different tests using the same criterion diagnosis can be compared to determine which scale is more diagnostically accurate (Hanley & McNeil, 1983). Furthermore, optimal cut points that compare the sensitivity and specificity can be estimated to maximize the diagnostic accuracy of a scale. To make these calculations easier for clinicians to use, diagnostic likelihood ratios (DLRs), which quantify the increased likelihood of having the disorder in question based on the individual's scale score, can be calculated. Probability nomograms or calculators combine base rate probabilities with multiple sources of information, such as likelihood ratios from scales or risk factors, to determine a posterior probability that an individual meets criteria for a particular disorder. This diagnostic approach improves accuracy of assessment and diagnosis, as well as streamlines successful treatment by reducing clinician bias and incorporating multiple data sources (Jenkins, Youngstrom, Washburn, & Youngstrom, 2011; Youngstrom, Choukas-Bradley, Calhoun, & Jensen-Doss, 2015).

Why Focus on the Child and Adolescent Symptom Inventory?

There are hundreds of measures for depression (Klein, Dougherty, & Olino, 2005; Nezu, Ronan, Meadows, & McClure, 2000). However, a much smaller number of these measures

have been examined using diagnostic efficiency statistics to determine their sensitivity and specificity in identifying depressive disorders (for a review, see Stockings et al., 2015).

Among the various screening tools available is the Child and Adolescent Symptom Inventory (CASI-4R; Gadow & Sprafkin, 2005), a youth (ages 5 to 18) rating scale for affective and behavior symptoms. It is one of the few that actually covers all of the DSM symptoms of depression; other widely used measures either predated the DSM-III-R or DSM-IV when their item pools were developed (Achenbach & Rescorla, 2001) or the scale is too short to provide complete coverage (see Stockings et al., 2015). Moreover, the CASI-4R assesses multiple other childhood disorders in addition to depression and dysthymia (ADHD, ODD, conduct disorder, generalized anxiety disorder, social phobia, separation anxiety disorder, bipolar disorder, schizophrenia, autism and Asperger's disorder, anorexia, and bulimia). This wider breadth of coverage allows for the simultaneous assessment of other common disorders, unlike other scales that only assess only one symptom domain (e.g., Child Depression Inventory, Mood and Feelings Questionnaire; Costello & Angold, 1988; Kovacs, 1992). The CASI-4R combines items from the Child Symptom Inventory (CSI-4; Gadow & Sprafkin, 2002) and Adolescent Symptom Inventory (ASI-4; Gadow & Sprafkin, 1998). It contains a major depressive episode (MDE) and dysthymic disorder subscale to examine youth depressive symptoms. The CASI-4R includes parent and teacher reports, takes 15 to 20 minutes to administer (all subscales), and uses both a dimensional (severity) and categorical (present/absent) assessment of depressive symptoms. Given its ease of use and correspondence with DSM-IV criteria, the CASI-4R has the potential to be a valuable screening tool for youth depressive symptoms. The severity scale also could add value as a measure of initial status and treatment response.

Prior research with the CASI-4R

Several studies show that the CASI-4R exhibits satisfactory internal consistency, test-retest-reliability, as well as convergent and divergent validity from various assessment measures in both community-based and clinic samples (reviewed by Gadow & Sprafkin, 2015). In regard to the depression and dysthymia subscale, preliminary research with the CSI-4 has found that the internal consistency is moderate for the parent version and high for the teacher version (Gadow & Sprafkin, 2002). Convergent and discriminant validity have been examined comparing the CSI-4 parent and teacher reports to the Child Behavior Checklist (CBCL) and Teacher Report Forms (TRF), and the depression scales significantly correlated with the Withdrawn, Somatic Complaints, and Anxious/Depressed subscales of the CBCL and TRF, and the CASI-4R dysthymia portion correlated with the CBCL Anxious/Depressed subscales (Gadow & Sprafkin, 2002).

However, less is known about the clinical utility of the CASI-4R for depressive disorders in youth. One study examined the clinical utility of the CSI-4 parent version in a sample of 247 clinically referred boys between the ages of six and 11; results showed that internal consistency and test-retest reliabilities were comparable to other behavior rating scales, and its convergent validity was established with the CBCL and the Diagnostic Interview for Children and Adolescents – Revised Parent Version (DICA-P; Reich, Shayka, & Taibelson, 1991; Sprafkin, Gadow, Salisbury, Schneider, & Loney, 2002). Furthermore, for moderate

symptom severity scores, the CSI-4 demonstrated high sensitivity (0.90), lower specificity (0.47), and percentage-correct classification scores were around 89% for screening cutoff scores for depression diagnoses (as assessed by the DICA-P; Sprafkin et al., 2002). These findings suggest that the CSI-4 is useful as a screening instrument for depressive disorders in clinically referred boys with behavioral and emotional problems. However, it is important to consider that the diagnosis was determined solely by parent report, so there could be an artificially inflated association between the diagnosis and the predictor measure, which is also a parent report. Furthermore, more research is needed to determine how the CASI-4R (which includes both the CSI-4 and ASI-4) compares in its diagnostic performance for both males and females.

Current Study

Although several studies have examined validity and reliability of the CASI-4R depressive and dysthymic disorder subscale for youth mood disorders, less is known about how an individual's subscale score correspond to the likelihood of having a mood disorder. Furthermore, previous research has focused on community samples, primary care populations, or on one sex (e.g., males; Sprafkin et al., 2002), which limits the generalizability and clinical utility of the results. Therefore, the present study examined youth ages 6 to 12 seeking services in an outpatient mental health care setting. We focused on depression in a sample with high rates of externalizing disorders and mood dysregulation, and particularly mania, for several reasons. Externalizing issues are a more common driver of outpatient referrals, but frequently mood plays a core role in their etiology or maintenance (Stringaris, Cohen, Pine, & Leibenluft, 2009; Stringaris, Maughan, & Goodman, 2010). In addition, depression has a complex longitudinal course, with externalizing symptoms leading to later depression as well as a homotypic continuity pathway (Copeland, Shanahan, Egger, Angold, & Costello, 2014). Furthermore, early onset of mood symptoms may be associated with a bipolar disorder course (Mesman et al., 2016). Moreover, some data suggest that the first clinical episode of bipolar disorder may often be depression. This study had several aims. **Aim 1** examined diagnostic and clinical utility of the CASI-4R depressive subscale for detecting youth mood disorders. We hypothesized that both scoring methods (symptom count, symptom severity) would significantly discriminate between youth with current mood disorders and youth with other diagnoses. **Aim 2** compared diagnostic efficiency of the symptom severity and symptom count scoring methods. We expected that with the number of items in each scoring form being equal, the symptom severity scoring would outperform the symptom count scoring because the severity scoring provides more information about the nature and extent of the depressive symptoms beyond noting the presence or absence of a particular symptom. **Aim 3** compared diagnostic efficiency of the CASI-4R parent and teacher reports. We hypothesized that parent reports would significantly outperform teacher reports, in line with previous research findings with other measures (Youngstrom et al., 2004). **Aim 4** examined sex and age differences in diagnostic efficiency of the CASI-4R subscale. We expected that sex and age could moderate effects of CASI scores on mood diagnosis, although the direction of moderation is ambiguous based on previous results of rates of endorsement of different symptoms such as crying and worrying about the future (van Beek, Hessen, Hutteman, Verhulp, & van Leuven, 2012). We

explored race as a covariate in predicting mood diagnosis. We also illustrated clinical utility of the CASI-4R subscale through the use of a clinical vignette.

Methods

Participants and Procedures

These secondary analyses used baseline data from the Longitudinal Assessment of Manic Symptoms (LAMS) study (NIH R01MHREDACTED for masked review), which examined the trajectory of children and youth with elevated symptoms of mania (ESM) seeking treatment in nine outpatient mental health clinics. The Institutional Review Boards at the four university-affiliated LAMS sites (REDACTED for masked review) approved all study procedures.

Parents/guardians of children between the ages of 6 and 12.9 years visiting one of nine child outpatient mental health clinics from four university-affiliated LAMS sites were asked to complete the Parent General Behavior Inventory-10-Item Mania Scale (PGBI-10M) (Youngstrom, Frazier, Demeter, Calabrese, & Findling, 2008; Youngstrom et al., 2005) to assess their children for the presence of ESM in the past six months. Children whose parents/guardians rated them at or above a score of 12 (ESM+) on the PGBI-10M were invited to participate in the longitudinal section of the LAMS study. A smaller subset of children scoring <12 on the PGBI-10-M (ESM-), matched by age, sex, ethnicity, and Medicaid status, were also invited to participate as a comparison group. Eligible participants in the main outcome study must (1) not have received mental health treatment in one of the LAMS outpatient clinic sites in the past 12 months; (2) speak English; (3) have a parent/guardian who speaks English; (4) not have a sibling in the same household who had previously participated in the LAMS screening (see Horwitz et al., 2010 for additional details). Of the 1,124 children who met the score threshold for ESM+ and 1,498 in the ESM- (396 of those were invited to serve as controls), 621 from ESM+ and 86 ESM- participants enrolled in the study. Out of the participants enrolled in the study, 7 were subsequently determined to be ineligible, so 700 participants were included in the study, of whom 685 had complete data on the variables analyzed here.

Measures

Diagnoses—To assess for DSM-IV diagnostic criteria, youth participants and their guardians completed the Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Episode (K-SADS-PL; Kaufman et al., 1997) along with the mood modules from the Washington University in St. Louis Kiddie Schedule for Affective Disorders (WASH-U K-SADS; Geller, Warner, Williams, & Zimmerman, 1998; Geller et al., 2001). Interviewer training consisted of attending a 3-day start-up meeting, watching and rating taped K-SADS-PL interviews, and leading administrations of assessment instructions. Trained research assistants interviewed youth participants and their caregivers, and a licensed clinical psychologist or psychiatrist reviewed K-SADS-PL data and diagnoses. Interrater reliability for the K-SADS-PL psychiatric diagnoses reached a kappa of 0.82 (Findling et al., 2010). CASI-4R scores were not available to either assessors or diagnosticians at consensus reviews.

Current depressive mood state was defined using a priori thresholds (Duax, Youngstrom, Calabrese, & Findling, 2007) from the Children's Depression Rating Scale-Revised (Poznanski & Mokros, 1996). Using diagnoses derived from the K-SADS, we created three categories of participants with current depressive symptoms:

Narrow definition: Children were included if they met criteria for one of the following disorders: major depressive disorder (also including single episode, recurrent, with psychotic features), dysthymic disorder (now called persistent depressive disorder in DSM-5; American Psychiatric Association, 2013), bipolar I or II disorder (most recent episode depressed), or schizoaffective disorder (depressive type). We treated the narrow definition as primary for analyses because any of these are likely to indicate a clinically significant mood disturbance related to depression, suggesting that if the CASI-4R scale identified a case, it would be a “true positive” in terms of requiring attention for the depressive symptoms.

Broad definition: This category included youth that met criteria for any diagnosis in the narrow category or any of the following: depressive disorder NOS, subsyndromal depression, bipolar disorder (unspecified subtype), bipolar I disorder (also including most recent episode hypomanic, manic, mixed, or unspecified), bipolar II disorder, cyclothymic disorder, bipolar disorder NOS, mood disorder NOS, premenstrual dysphoric disorder (PMDD), schizoaffective disorder (bipolar type), and adjustment disorder (with depressed mood, with mixed anxiety and depressed mood).

Very broad definition: This included children with any diagnosis from the narrow or broad categories, or any of the following: bipolar I disorder (single manic episode), mood disorder due to a general medical condition, substance-induced mood disorder, subsyndromal bipolar disorder, and adjustment disorder with mixed disturbance of emotions and conduct.

Comparison group: The comparison group comprised those who did not meet criteria for the given target diagnosis category, so the number in the comparison group changed depending on the diagnostic category being analyzed. As all participants were treatment-seeking individuals, there were no healthy controls in the comparison group.

Child and Adolescent Symptom Inventory – 4R (CASI-4R; (Gadow & Sprafkin, 2005). The CASI-4R (163-item parent report, 120- item teacher report) examines presence of DSM-IV symptoms for emotional and behavioral disorders in children and adolescents. These analyses examined parent- and teacher-reported scores for the major depressive episode (MDE) and dysthymic disorder subscale. The MDE and dysthymic disorder subscale contains 8 items describing symptoms related to DSM-IV criteria. Parents and teachers rate symptom frequency on a 0 to 3 scale (0 – *Never*, 1 – *Sometimes*, 2 – *Often*, 3 – *Very Often*). Sample items include, “[Youth] is depressed for most of the day,” and “[Youth] has little confidence, feels inferior to others, or is very self-conscious.” In addition, informants answer *yes* or *no* to an additional seven items related to the child's other depression related symptoms. Sample items included, “[Youth] has experienced a big change in his/her normal appetite or weight,” and “[Youth] has experienced a big change in his/her ability to concentrate or make decisions.”

We compared two scoring methods to determine which scoring system worked best at predicting a depressed mood diagnosis, symptom severity and symptom count scoring:

Symptom severity scoring summed the seven DSM items scored 0-3. These items are: “Is depressed for most of the day,” “Shows little interest in (or enjoyment of) pleasurable activities,” “Has low energy or is tired for no apparent reason,” “Feels worthless or guilty,” “Has little confidence, feels inferior to others, or is very self-conscious,” “Talks about death or suicide,” and “Feels that things never work out right.” Next, the following four *yes* or *no* items were recoded (*no* - 0.5 and *yes* - 2.5, following the published scoring criteria) and added to the DSM items: “Has experienced a big change in his/her normal appetite or weight,” “Has experienced a big change in his/her normal sleeping habits (trouble sleeping or sleeps too much),” “Has experienced a big change in his/her normal activity level (overactive or inactive),” and “Has experienced a big change in his/her ability to concentrate or make decisions.”

Symptom count scoring recoded the above seven DSM items so that scores of 0 and 1 were recoded to 0 and scores of 2 or 3 were recoded to 1. They are then then summed with the four *yes* or *no* items (where 0 – *no* and 1 – *yes*).

Both scores excluded four items that were not symptoms of depression, *per se*, but rather asked about impairment, appearing more sensitive or tearful than usual, or the presence of a precipitating event. These items are intended to help the clinician and are not included as a part of the diagnostic algorithm.

Informants: Both the caregiver and teacher reports include symptom severity and symptom count scores. The number of items across informants is similar with the exception that the teacher report (both forms) excludes the two items related to changes in eating and sleeping habits.

Data Analytic Plan

Analyses used SPSS (Version 22.0) and the pROC (Robin et al., 2011) and Aod packages in R. First, we examined baseline demographic and clinical characteristics, checking for potential outliers. All tests used a two-tailed alpha-level of .05. Using independent samples *t*-tests, Levene's *F* tests of variances, or chi-square tests, we compared clinical and demographic characteristics (age, race, sex, number of comorbid diagnoses, CASI-4R scores) between children in the target diagnosis category (narrow for the primary analyses; broad or very broad as supplemental) and children not in that category.

Back-to-back histograms for all distributions looked for any degenerate distributions or outliers. Next, nonparametric estimates of the AUC from ROC analyses (Youngstrom, 2014) quantified diagnostic efficiency of the CASI-4R depression symptom severity and symptom count scoring forms (Aim 1). We calculated AUCs for the three diagnostic target categories (narrow, broad, and very broad) using both CASI-4R scoring methods to evaluate which scoring method better discriminates youth with depressive symptoms from other youth in the sample (Aim 2). In addition, we calculated AUCs using both caregiver and teacher versions of the scales (Aim 3). AUC guidelines suggest that values in the mid .50s are considered

small (not clinically useful), mid .60s are medium (may provide incremental clinical value), low to mid .70s are large (clinically informative), low .80s are excellent, and high .80s are considered exceptional under clinically rigorous designs (highly informative) (Rice & Harris, 2005). Values higher than .90 in mental health research often indicate a design flaw, such as using a comparison group that is too distinct or that excludes common conditions that might create false positives (e.g., healthy controls; Youngstrom, Genzlinger, Egerton, & Van Meter, 2015; Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006).

Next, permutation tests (Venkatraman, 2000) compared ROC curves between CASI-4R scoring forms, as well as teacher and caregiver forms for each target diagnostic category (Aims 2 and 3). This test compared ROC curves at all the operating points instead of just the overall AUC estimates, which gives the test more statistical power to detect significant differences even when AUCs appear the same at the second decimal point (Venkatraman, 2000). For CASI-4R scales with the highest diagnostic efficiency (caregiver or teacher version, symptom severity and symptom count scoring versions), hierarchical logistic regressions tested whether these scales still significantly predicted target diagnoses even after controlling for demographics (sex, age, race) and number of comorbidities. Model 1 included demographic and clinical variables, as well as CASI-4R scores as predictors of mood diagnosis. Model 2 tested whether sex and age moderated accuracy of the scale by calculating sex by scale and age by scale interaction terms (Aim 4).

Next, we calculated DLRs from optimal cut-points to produce the best combination of sensitivity and specificity from ROC curves (Robin et al., 2011). The formula is as follows:

$$DLRs = \frac{\text{True Positive Rate}/(\text{True Positive Rate} + \text{False Negative Rate})}{\text{False Positive Rate}/(\text{False Positive Rate} + \text{True Negative Rate})}$$

DLRs provide clinically useful information regarding odds of a diagnosis associated with a particular test score. DLRs less than 1.0 are associated with lower than average odds, at or around 1.0 indicate no change in odds, 2-5 indicate a small increase in the odds and possibly clinically significant, 5-10 are a large increase, and DLRs greater than 10 are odds changes that are likely clinically decisive (Straus, Glasziou, Richardson, & Haynes, 2011).

Consistent with prior work (Ong et al., 2016; Van Meter et al., 2016) and commonplace in epidemiological work to get more accurate population parameter estimates (Little, Lewitzky, Heeringa, Lepkowski, & Kessler, 1997), we weighted the data used in these analyses to address the unequal probability of selection and sample response. First, a base weight was created by utilizing the inverse of the probability of selection. Next, logistic regressions estimated the probability of participation in the LAMS study, including age, sex, Hispanic ethnicity, insurance status, the main effect of site, ESM status, and all two-way interactions with ESM (except for Hispanic ethnicity because of its small cell size) as predictors in the final model. Ten propensity cells were created from the distribution of predicted probabilities. The non-response weight was obtained using the inverse of the unweighted response rate per propensity cell (Little & Vartivarian, 2003). The resulting final weight was calculated by taking the product of the base weight and non-response weight, rescaled so

that the sum was equal to the sample size. This weighting process created a sample that more closely matches the overall presenting clinical population, reducing the risk of misleading results, and has been used in previous studies using similar analytic methods (Ong et al., 2016; Van Meter et al., 2016).

Results

Tables and results explained in this section primarily refer to the narrow target diagnosis category. Supplemental tables display results for the broad and very broad diagnosis categories.

Demographics and Preliminary Analyses

Out of the 700 participants, 66 met criteria for a narrow mood diagnosis, 259 for a broad mood diagnosis, and 269 for the very broad mood diagnosis category. Within the narrow mood diagnosis, 63 of the participants who met criteria for a narrow mood diagnosis and 622 who did not meet criteria for a mood diagnosis had a completed CASI-4R caregiver form. In contrast, 42 of the participants who met criteria for a narrow mood diagnosis and 418 who did not meet criteria had a completed CASI-4R teacher form (460 total). The lower numbers of completed teacher forms are the result of not all teachers returning the form, and this rate of nonresponse is similar to other studies not taking place in a school (Youngstrom, Joseph, & Greene, 2008). Participants without teacher data did not differ in age, sex, number of diagnoses at baseline, or CASI-4R caregiver scores ($ps > .08$). Participants without teacher data were significantly less likely to be white (69% versus 56%, $p < .001$). CASI-4R depression scores were calculated and diagnostic efficiency statistics were computed for participants whose caregivers completed at least 9 of the 11 required items or teachers who completed at least 7 of the 9 required items. We prorated caregiver and teacher scores if respondents skipped one or two items, so it is possible to observe non-integer scores in these analyses. Parent report was 99.96% complete and teacher report was 98.75% complete at the item level.

Table 1 presents the demographic and clinical differences between children with any narrow mood diagnosis versus no narrow mood diagnosis. Children in the narrow mood group were significantly older ($p < .001$) and had more Axis I diagnoses ($p < .001$). Therefore, logistic regressions examined using these as covariates or potential moderators of CASI-4R score accuracy. There were no group differences in race ($p > .05$). The narrow mood group also had higher CASI-4R caregiver and teacher scores (both scoring methods; $ps < .05$). This is preliminary evidence that the CASI-4R depression and dysthymia subscale has discriminative validity.

Diagnostic Efficiency Analyses

Table 2 presents AUCs from ROC analyses for the CASI-4R caregiver and teacher reports of symptom severity and symptom count scoring forms. Back-to-back histograms for all distributions showed there were no degenerate distributions or problematic outliers (Youngstrom, 2014). AUCs for caregiver reports were large and clinically informative (symptom severity AUC = .79, symptom count AUC = .78, $p < .001$). Venkatraman's test

compared AUCs for the symptom severity versus symptom count scoring forms. The caregiver symptom severity scoring form outperformed the symptom count scoring form ($p=.001$), suggesting a small but reliable advantage for detecting the presence of a narrowly defined depressive disorder. In contrast, teacher symptom severity did poorly ($AUC = .56, p > .05$). The teacher symptom count report was statistically significant but corresponds to a small to medium effect size ($AUC = .60$). See Figure 1 for ROC curves.

When re-running the analyses using a broad definition of mood disorder, the prevalence increased to 37%, and the AUCs and the DLRs were similar. The caregiver symptom severity scores ($AUC = .81$) were significantly better than the symptom count scores ($AUC = .77; p < .001$), and both teacher forms were statistically significant ($ps < .05$), both corresponding to small effect sizes. Like with the narrow mood, the teacher symptom count scoring ($AUC = .56$) outperformed the symptom severity scoring ($AUC = .55; p < .001$). When using the very broad definition of a mood disorder, the prevalence increased to 38%, and the AUCs and DLRs were similar (see supplemental tables). Like with the broad mood definition, all scoring forms were statistically significant, and the caregiver symptom severity scoring ($AUC = .81$) outperformed the symptom count scoring ($AUC = .77; p < .001$). Both teacher forms corresponded to a small effect size, and the teacher symptom count form ($AUC = .56$) slightly outperformed the teacher symptom severity form ($AUC = .55, p < .001$).

The optimal cut scores for the broad and very broad mood disorder groups were at 7.50, similar to the narrow mood diagnosis. As expected, when going from narrow to broad and very broad the sensitivity decreased at the optimal cut score – more cases with milder disorders counted as targets, and sometimes obtained low scores on the CASI-4R, and specificity increased slightly – some cases with high CASI-4R scores were reclassified as “target” cases, even though not meeting formal criteria for a major depressive disorder or similar narrowly defined diagnosis.

Covariate and Moderator Analyses

Logistic regressions examined whether the best-performing scale, the caregiver symptom severity scoring form, still significantly predicted the narrow mood target diagnosis even after controlling for demographics (sex, age, race) and number of diagnoses. Model 1 controlled for demographic variables and number of diagnoses. CASI-4R scores were significant predictors of narrow mood diagnosis after controlling for demographics and number of diagnoses ($B = 0.13, SE = 0.02, z = 6.19, p < .001$). This overall model explained 26% of the variance in narrow mood diagnosis. Being white, age, and number of Axis I diagnoses were significant predictors of narrow mood diagnosis ($ps < .05$). Being female did not predict narrow mood diagnosis ($p = .32$).

Model 2 added sex by CASI-4R and age by CASI-4R interaction terms (Aim 4) to determine whether age or sex moderated effects of CASI-4R scores on narrow mood diagnosis. The age by CASI-4R score and sex by CASI-4R score interaction terms were not significant ($ps > .12$). A final supplemental analysis tested whether the CASI-4R depression scale showed incremental validity even after controlling for other scales on the CASI-4R; this provided a strong test of whether shared method variance was driving the results. The depression scale

continued to show a strong incremental weight ($p < .0005$) even controlling for the other scales.

Diagnostic Likelihood Ratios (DLRs)

We estimated DLRs using cut-off scores that optimized the combination of sensitivity and specificity. Table 3 reports DLRs for the CASI-4R depression symptom severity scoring. Using optimal cut scores, a CASI-4R depression severity score of 7.50 or higher (Sensitivity = 62%; Specificity = 89%) resulted in moderate increase in odds (over two times) of meeting criteria for a narrow mood diagnosis ($DLR = 2.31$). In contrast, scores below 7.50 were associated with reduced likelihood of receiving a narrow mood diagnosis ($DLR = 0.18$). To more closely examine the effects of race on CASI-4R scores, we compared optimal cut DLRs to see if they were different from one another. Optimal thresholds for whites versus nonwhite participants were four points apart: the cut point for white participants (AUC = .78) was 7.50, whereas for nonwhite participants (AUC = .83) it was 11.50. If replicated, then this suggests that using the same threshold might overestimate the probability of depression for some cases with moderate scores.

To provide more informative DLRs for clinicians to use (Straus et al., 2011), scores were divided into four strata (see Table 3). Youth who received a narrow mood diagnosis were almost 4 times more likely to score 15 or higher on the CASI-4R depression symptom severity subscale ($DLR = 3.73$). In contrast, individuals with CASI-4R depression severity scores below 8 had decreased odds of meeting criteria for a narrow mood diagnosis ($DLR = 0.18$).

Discussion

The overarching goal of the present study was to examine diagnostic efficiency and clinical utility of the CASI-4R depressive subscale for detecting youth mood disorders. We found that when using caregiver report, both symptom count and symptom severity scoring methods significantly predicted narrow mood diagnosis, but symptom severity performed statistically significantly better than the symptom count scoring method. This finding is consistent with our hypothesis that the symptom severity scoring method would perform better due the psychometric advantages of retaining more information about the variability of symptom presentations. However, the difference in AUCs between these two scoring methods was .01, so the difference is unlikely to be clinically meaningful. A pragmatic benefit to using the severity scoring method over the symptom count method is it eliminates the need to recode all of the items (only recoding four items), which leads to ease of use and application in a clinical setting. Findings were similar when using the broad and very broad mood definitions; although statistically significant differences were observed between the severity and symptom count scoring methods, the differences were small and not likely to be clinically meaningful.

We also found that the CASI-4R teacher symptom severity report did not perform above chance at predicting a narrow mood diagnosis, and the teacher symptom count report corresponded to a small effect size. These findings are consistent with our hypothesis, as well as in line with results of previous studies (Jensen et al., 1999; Youngstrom et al., 2004;

Youngstrom, Genzlinger, et al., 2015). Because the CASI-4R asks the informant to report on the child's mental state (e.g., feeling worthless/guilty, talking about death/suicide, feels inferior to others), as well as behaviors that may be more commonly noticed outside the school setting (e.g., appetite and sleep disturbance, which are omitted from the teacher form), teachers in their limited time with a student may not note changes in these areas. It is possible that parent report had an advantage in that parents were also involved in the diagnostic interview, and teachers were not. This concern is allayed somewhat by the fact that the depression score showed substantial incremental validity even after controlling for other CASI-4R scales, controlling for shared method variance. Furthermore, the diagnostic interviews all included the youth as well as the caregiver, adding mental status, behavioral observations, and youth report to the sources of diagnostic information.

The CASI-4R severity scores still significantly predicted narrow mood diagnosis after controlling for demographics and number of comorbidities. The age by CASI-4R or sex by CASI-4R interactions were not significant at predicting narrow mood diagnosis. We found that CASI-4R scores were significantly better at predicting narrow mood diagnosis for nonwhite participants, which means that if clinicians use the same threshold for all participants, they may overestimate the likelihood of depression for some youths who score in the moderate range. However, our study focused on diagnostic accuracy, so future research should examine whether differential item functioning is present in the CASI-4R depending on youth sex, age, and race.

The CASI-4R depression AUCs for predicting a narrow mood diagnosis were large and clinically informative. When comparing these results to AUCs for other scales, several other scales, such as the Child Depression Inventory (CDI; Kovacs, 1992), Mood and Feelings Questionnaire (MFQ; Costello & Angold, 1988), and the Beck Depression Inventory (BDI; Beck & Steer, 1993) produce higher AUCs (Stockings et al., 2015). However, many of these previous studies use different designs and often include community controls or matched healthy controls as their comparison group, which can inflate DLRs (Bossuyt et al., 2003). Therefore, for clinicians to be able to come to a more accurate assessment, it is important to use DLRs derived from similar settings (i.e., outpatient, inpatient, community mental health). The CDI is one of the few other measures that has been evaluated for distinguishing depressed youth from youth with anxiety disorders (Comer & Kendall, 2005), as well as referred youth with disorders other than depression (Timbremont, Braet, & Dreessen, 2004). A strength of our study was that we conducted analyses for data about youth recruited from outpatient clinical settings and used a comparison group composed of children who had a range of other disorders, making it more similar to what clinicians may experience in their practices. One important future direction of research should be to examine how the CASI-4R compares in diagnostic efficiency in clinical settings to other incumbent measures.

One potential obstacle clinicians may face when choosing between administering the CASI-4R and other scales is its cost. However, the CASI-4R is not solely a depression measure; it contains subscales for all the DSM-IV childhood disorders.

By including ROC analyses and DLRs in our analyses, it becomes easier for clinicians to apply actuarial methods in their assessment practices by incorporating information from

multiple sources (e.g., scales, family history) in a systematic way that improves prediction of diagnosis, while also limiting the role of bias and heuristics that can influence clinician judgment (Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Below, we illustrate the use of DLRs to show how multiple sources of information can be integrated in a systematic way.

Clinical Vignette

Jacob, a 10 year-old Hispanic male, is referred to the clinic by his mother, Celeste, due to complaints from the teacher about his behavior. Jacob's teacher reports that Jacob has become increasingly irritable in the classroom, often "talking back" to her when asked why he has not completed his homework assignments, which has increasingly been more frequent. Several of his classmates have made complaints to the teacher about Jacob consistently picking on them and calling them names. In addition, Jacob has been disruptive in the classroom, making loud comments such as, "This is stupid, why do we need to learn this?" as his teacher is giving a lesson. Moreover, Jacob's teacher describes his demeanor as increasingly "sullen" or sulky, and he gets distracted easily when he is supposed to be engaging in solitary activities, such as writing or doing math problems.

Celeste reports that she has noticed Jacob increasingly isolating himself and retreating to his room after school. They have begun arguing more as a result of Celeste checking with Jacob about completing his homework, which irritates Jacob to the point where he has yelled comments such as, "I will never be able to learn this because I am so stupid!" He has anger outbursts; he begins to scream and yell whenever Celeste tells him to do his chores or clean his room; she says that every disagreement they have seems to turn into a "screaming match." Initially, Celeste thought that these behaviors were due to typical parent-child disagreements, but now she has expressed concern because she thinks that Jacob seems "angry all the time."

While you are conducting a clinical interview, you learn that Celeste suffered from severe depression several months after immigrating to the U.S. about nine years ago. Celeste was treated with an SSRI for two years before she discontinued because she "was feeling better and no longer needed it." Among other measures, you ask Celeste to fill out the CASI-4R to assess Jacob for depression, and her score on the symptom severity subscale was a 17.

You can use a probability nomogram to integrate this information (see Figure 2). First, you start with the base rate of diagnosis for depression in outpatient settings similar to your clinic, which from prior research, you can say is 17% (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009), and you plot this value on the pretest probability line. Next, as you know Jacob has a family history of depression, that elevates his risk three times (Weissman et al., 2006). You plot this value (DLR = 3.0) on the likelihood ratio line, draw a line from the pretest probability value to the DLR, then extend the line to the posttest probability line. From there, you can see the posttest probability is about 38%. Next, you add a second nomogram and plot the 38% to the pretest probability line. You can then incorporate the DLR for Celeste's CASI-4R score (from Table 3, you see it is 3.73) by plotting it in the second line of the nomogram. Again, you draw a line from the 38% through 3.73 to get a posterior probability of approximately 70% (see Figure 2). If you collect additional

information, such as structured/semi-structured interviews or a self-report scale from Jacob, you can continue to integrate the DLRs and adjusting the posttest probability as shown above.

It is important to note that results of this nomogram are not meant to be considered as a diagnosis, but rather as a guide to inform the clinician as to what the next steps should be (Jenkins et al., 2011), whether it be formally confirming a diagnosis of a mood disorder through the use of a structured diagnostic interview; or if the posttest probability for depression was low, considering other potential disorders that may account for Jacob's behavior.

Strengths and Limitations

This paper is the first to report AUCs and DLRs for CASI-4R depressive and dysthymic disorder subscale in both girls and boys, as well as in an outpatient sample. This study also used a sample of children with a wide range of diagnoses and no healthy controls, making the AUCs and DLRs more generalizable to what clinicians may observe in their clinical setting. However, findings from this paper should be interpreted in the context of several limitations. First, the CASI-4R does not ask caregivers to report on their child's behavior during a specific time period, which makes it difficult to distinguish whether caregivers are reporting on current problems or past behaviors that are no longer an issue. In a clinical setting, clinicians may want to ask about a child's lifetime symptoms of depression as well as current symptoms to be able to distinguish what the current presenting issues are.

Second, no CASI-4R self-report scale was administered. Previous research suggests that caregivers are better reporters of a child's behaviors than children are of themselves, especially for externalizing behavior; however, for internalizing symptoms, given that caregivers may not be able to observe everything the child subjectively experiences, their reports may not be as accurate as self-report, particularly as the child gets older and more aware of his/her mental state. There are other specialized self-report scales that are developmentally appropriate and often are viable to use as self-report options with younger age ranges (Stockings et al., 2015). Third, given the primary objective of the original study was to track the progression of manic symptoms in children over time, ESM+ individuals comprise approximately 90% of the sample. To address this and make the findings more generalizable to other populations, we used weighting in our analyses, as is typically done in epidemiological studies. However, despite weighting procedures, our sample composition may have influenced the findings. Further research using the CASI-4R in other settings, such as private offices or inpatient units, may improve our confidence in its robustness as a depression assessment. Furthermore, much of the field is using DSM-5 criteria, and these findings are based on the DSM-IV diagnoses. Changes to depression criteria in the DSM-5 were minor (American Psychiatric Association, 2013) and unlikely to have a significant effect on the diagnostic accuracy of measures; however, this remains an empirical issue.

Conclusions

The caregiver CASI-4R depression and dysthymia subscale is useful for identifying youth mood disorders. Elevated symptom severity scores are related to a 2 to 4-fold increase in the

odds of meeting criteria for a narrow mood diagnosis. The CASI-4R is easy to administer, takes minutes to complete, and assesses a wide range of domains in addition to depressive symptoms. Having likelihood ratios for scales such as the CASI-4R allows the clinician to be able to determine the next course of action to take with a client, whether it is conducting follow-up assessments, and/or starting treatment immediately (Youngstrom, Choukas-Bradley, et al., 2015).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Achenbach TM, Rescorla LA. Manual for the ASEBA School-Age Forms & Profiles. Burlington, VT: University of Vermont; 2001.
- Ægisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, et al. Cohen G. The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*. 2006; 34(3):341–382. DOI: 10.1177/0011000005285875
- American Psychiatric Association. The Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 5th. Arlington, VA: American Psychiatric Publishing, Inc; 2013.
- Angold A, Costello EJ, Erkanli A. Comorbidity. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. 1999; 40(1):57–87. DOI: 10.1111/1469-7610.00424
- Bayes T, Price R. An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London (1683-1775)*. 1763; 53:370–418.
- Beck AT, Steer R. Beck Depression Inventory. San Antonio, TX: Psychological Corporation; 1993.
- Birmaher B, Arbelaez C, Brent D. Course and outcome of child and adolescent major depressive disorder. *Child and Adolescent Psychiatric Clinics of North America*. 2002; 11(3):619–637. x. [PubMed: 12222086]
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Group, S. Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Academic Radiology*. 2003; 10(6):664–669. [PubMed: 12809421]
- Comer JS, Kendall PC. High-end specificity of the children's depression inventory in a sample of anxiety-disordered youth. *Depression and Anxiety*. 2005; 22(1):11–19. DOI: 10.1002/da.20059 [PubMed: 15965988]
- Copeland WE, Shanahan L, Egger H, Angold A, Costello EJ. Adult Diagnostic and Functional Outcomes of DSM-5 Disruptive Mood Dysregulation Disorder. *American Journal of Psychiatry*. 2014; 171(6):668–674. DOI: 10.1176/appi.ajp.2014.13091213 [PubMed: 24781389]
- Costello EJ, Angold A. Scales to assess child and adolescent depression: checklists, screens, and nets. *Journal of the American Academy of Child and Adolescent Psychiatry*. 1988; 27(6):726–737. DOI: 10.1097/00004583-198811000-00011 [PubMed: 3058677]
- De Los Reyes A, Augenstein TM, Wang M, Thomas SA, Drabick DA, Burgers DE, Rabinowitz J. The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*. 2015; 141(4):858–900. DOI: 10.1037/a0038498 [PubMed: 25915035]
- De Los Reyes A, Kazdin AE. Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*. 2005; 131(4):483–509. DOI: 10.1037/0033-2909.131.4.483 [PubMed: 16060799]
- De Los Reyes A, Youngstrom EA, Swan AJ, Youngstrom JK, Feeny NC, Findling RL. Informant discrepancies in clinical reports of youths and interviewers' impressions of the reliability of

- informants. *Journal of Child and Adolescent Psychopharmacology*. 2011; 21(5):417–424. DOI: 10.1089/cap.2011.0011 [PubMed: 22040187]
- Duax JM, Youngstrom EA, Calabrese JR, Findling RL. Sex differences in pediatric bipolar disorder. *Journal of Clinical Psychiatry*. 2007; 68(10):1565–1573. [PubMed: 17960974]
- Dunn V, Goodyer IM. Longitudinal investigation into childhood- and adolescence-onset depression: psychiatric outcome in early adulthood. *British Journal of Psychiatry*. 2006; 188:216–222. DOI: 10.1192/bjp.188.3.216 [PubMed: 16507961]
- Edelbrock C, Costello AJ, Dulcan MK, Kalas R, Conover NC. Age differences in the reliability of the psychiatric interview of the child. *Child Development*. 1985; 56(1):265–275. [PubMed: 3987406]
- Essau CA. Course and outcome of major depressive disorder in non-referred adolescents. *Journal of Affective Disorders*. 2007; 99(1-3):191–201. DOI: 10.1016/j.jad.2006.09.010 [PubMed: 17049997]
- Findling RL, Youngstrom EA, Fristad MA, Birmaher B, Kowatch RA, Arnold LE, et al. Horwitz SM. Characteristics of children with elevated symptoms of mania: the Longitudinal Assessment of Manic Symptoms (LAMS) study. *Journal of Clinical Psychiatry*. 2010; 71(12):1664–1672. DOI: 10.4088/JCP.09m05859yel [PubMed: 21034685]
- Gadow KD, Sprafkin J. Child and adolescent symptom inventory-4R. Stony Brook, NY: Checkmate Plus; 2005.
- Gadow KD, Sprafkin J. The Symptom Inventories: An annotated bibliography. Stony Brook, NY: Checkmate Plus; 2015.
- Gadow KD, Sprafkin JN. Adolescent Symptom Inventory-4. Checkmate Plus; 1998.
- Gadow KD, Sprafkin JN. Child symptom inventory 4: Screening and norms manual. Checkmate Plus; 2002.
- Garber J, Horowitz JL. Depression in children. In: Gotlib IH, Hammen CL, Gotlib IH, Hammen CL, editors *Handbook of depression*. New York, NY: Guilford Press; 2002. 510–540.
- Geller B, Warner K, Williams M, Zimmerman B. Prepubertal and young adolescent bipolarity versus ADHD: assessment and validity using the WASH-U-KSADS, CBCL and TRF. *Journal of Affective Disorders*. 1998; 51(2):93–100. [PubMed: 10743842]
- Geller B, Zimmerman B, Williams M, Bolhofner K, Craney JL, DelBello MP, Soutullo C. 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 and Adolescent Psychiatry*. 2001; 40(4):450–455. DOI: 10.1097/00004583-200104000-00014 [PubMed: 11314571]
- Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*. 2000; 12(1):19–30. [PubMed: 10752360]
- Hammen C, Rudolph KD. Childhood mood disorders. In: Mash EJ, Barkley RA, editors *Child psychopathology*. New York, NY: The Guilford Press; 2003. 233–278.
- Hankin BL, Abramson LY, Moffitt TE, Silva PA, McGee R, Angell KE. Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology*. 1998; 107(1):128–140. [PubMed: 9505045]
- Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983; 148(3):839–843. DOI: 10.1148/radiology.148.3.6878708 [PubMed: 6878708]
- Horwitz SM, Demeter CA, Pagano ME, Youngstrom EA, Fristad MA, Arnold LE, et al. Findling RL. Longitudinal Assessment of Manic Symptoms (LAMS) study: background, design, and initial screening results. *Journal of Clinical Psychiatry*. 2010; 71(11):1511–1517. DOI: 10.4088/JCP.09m05835yel [PubMed: 21034684]
- Jenkins MM, Youngstrom EA, Washburn JJ, Youngstrom JK. Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice*. 2011; 42(2):121–129. DOI: 10.1037/a0022506
- Jensen PS, Rubio-Stipec M, Canino G, Bird HR, Dulcan MK, Schwab-Stone ME, Lahey BB. Parent and child contributions to diagnosis of mental disorder: are both informants always necessary? *Journal of the American Academy of Child and Adolescent Psychiatry*. 1999; 38(12):1569–1579. DOI: 10.1097/00004583-199912000-00019 [PubMed: 10596258]

- Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, et al. Ryan N. 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 and Adolescent Psychiatry*. 1997; 36(7):980–988. DOI: 10.1097/00004583-199707000-00021 [PubMed: 9204677]
- Kessler RC, Avenevoli S, Costello EJ, Georgiades K, Green JG, Gruber MJ, et al. Merikangas KR. Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. *Archives of General Psychiatry*. 2012; 69(4):372–380. DOI: 10.1001/archgenpsychiatry.2011.160 [PubMed: 22147808]
- Kessler RC, Avenevoli S, Ries Merikangas K. Mood disorders in children and adolescents: an epidemiologic perspective. *Biological Psychiatry*. 2001; 49(12):1002–1014. [PubMed: 11430842]
- Kessler RC, Merikangas KR, Wang PS. Prevalence, comorbidity, and service utilization for mood disorders in the United States at the beginning of the twenty-first century. *Annual Review of Clinical Psychology*. 2007; 3:137–158. DOI: 10.1146/annurev.clinpsy.3.022806.091444
- Klein DN, Dougherty LR, Olino TM. Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*. 2005; 34(3): 412–432. DOI: 10.1207/s15374424jccp3403_3 [PubMed: 16026212]
- Kovacs M. *Children's Depression Inventory: Manual*. North Tonawanda, NY: Multi-Health Systems; 1992.
- Kovacs M. Presentation and course of major depressive disorder during childhood and later years of the life span. *Journal of the American Academy of Child and Adolescent Psychiatry*. 1996; 35(6): 705–715. DOI: 10.1097/00004583-199606000-00010 [PubMed: 8682751]
- Kruschke JK. *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York, NY: Academic Press; 2011.
- Lewinsohn PM, Essau CA. Depression in adolescents. In: Gotlib IH, Hammen CL, Gotlib IH, Hammen CL, editors *Handbook of depression*. New York, NY, US: Guilford Press; 2002. 541–559.
- Lewinsohn PM, Hops H, Roberts RE, Seeley JR, Andrews JA. Adolescent psychopathology: I. Prevalence and incidence of depression and other DSM-III-R disorders in high school students. *Journal of Abnormal Psychology*. 1993; 102(1):133–144. [PubMed: 8436689]
- Lewinsohn PM, Rohde P, Seeley JR, Klein DN, Gotlib IH. Psychosocial functioning of young adults who have experienced and recovered from major depressive disorder during adolescence. *Journal of Abnormal Psychology*. 2003; 112(3):353–363. [PubMed: 12943014]
- Little RJ, Lewitzky S, Heeringa S, Lepkowski J, Kessler RC. Assessment of weighting methodology for the National Comorbidity Survey. *American Journal of Epidemiology*. 1997; 146(5):439–449. [PubMed: 9290504]
- Little RJ, Vartivarian S. On weighting the rates in non-response weights. *Statistics in Medicine*. 2003; 22(9):1589–1599. [PubMed: 12704617]
- McFall RM, Treat TA. Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology*. 1999; 50:215–241. DOI: 10.1146/annurev.psych.50.1.215
- Mesman E, Birmaher BB, Goldstein BI, Goldstein T, Derks EM, Vleeschouwer M, et al. Hillegers MH. Categorical and dimensional psychopathology in Dutch and US offspring of parents with bipolar disorder: A preliminary cross-national comparison. *Journal of Affective Disorders*. 2016; 205:95–102. DOI: 10.1016/j.jad.2016.06.011 [PubMed: 27423424]
- Nezu AM, Ronan GF, Meadows EA, McClure KS. *Practitioner's guide to empirically based measures of depression*. Dordrecht, Netherlands: Kluwer Academic Publishers; 2000.
- Ong ML, Youngstrom EA, Chua JJ, Halverson TF, Horwitz SM, Storfer-Isser A, et al. Group, L. Comparing the CASI-4R and the PGBI-10 M for Differentiating Bipolar Spectrum Disorders from Other Outpatient Diagnoses in Youth. *Journal of Abnormal Child Psychology*. 2016; doi: 10.1007/s10802-016-0182-4
- Pettit JW, Hartley C, Lewinsohn PM, Seeley JR, Klein DN. Is liability to recurrent major depressive disorder present before first episode onset in adolescence or acquired after the initial episode? *Journal of Abnormal Psychology*. 2013; 122(2):353–358. DOI: 10.1037/a0032655 [PubMed: 23713498]
- Poznanski EO, Mokros HB. *Children's depression rating scale, revised (CDRS-R)*. Western Psychological Services; Los Angeles: 1996.

- Rao U, Ryan ND, Birmaher B, Dahl RE, Williamson DE, Kaufman J, et al. Nelson B. Unipolar depression in adolescents: clinical outcome in adulthood. *Journal of the American Academy of Child and Adolescent Psychiatry*. 1995; 34(5):566–578. DOI: 10.1097/00004583-199505000-00009 [PubMed: 7775352]
- Reich W, Shayka T, Taibelson C. *The Diagnostic Interview for Children and Adolescents - Revised*. St. Louis, MO: Washington University; 1991.
- Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research*. 2009; 18(3):169–184. DOI: 10.1002/mpr.289 [PubMed: 19701924]
- Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*. 2005; 29(5):615–620. DOI: 10.1007/s10979-005-6832-7 [PubMed: 16254746]
- Richters J, Pellegrini DS. Depressed mothers' judgments about their children: An examination of the depression-distortion hypothesis. *Child Development*. 1989; 60(5):1068–1075. DOI: 10.2307/1130780 [PubMed: 2805884]
- Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011; 12(1):77. [PubMed: 21414208]
- Sprafkin J, Gadow KD, Salisbury H, Schneider J, Loney J. Further evidence of reliability and validity of the Child Symptom Inventory-4: parent checklist in clinically referred boys. *Journal of Clinical Child and Adolescent Psychology*. 2002; 31(4):513–524. DOI: 10.1207/S15374424JCCP3104_10 [PubMed: 12402570]
- Stockings E, Degenhardt L, Lee YY, Mihalopoulos C, Liu A, Hobbs M, Patton G. Symptom screening scales for detecting major depressive disorder in children and adolescents: a systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders*. 2015; 174:447–463. DOI: 10.1016/j.jad.2014.11.061 [PubMed: 25553406]
- Straus SE, Glasziou P, Richardson WS, Haynes RB. *Evidence-based medicine: How to practice and teach it*. 4. Churchill Livingstone; Elsevier: 2011.
- Stringaris A, Cohen P, Pine DS, Leibenluft E. Adult outcomes of youth irritability: A 20-year prospective community-based study. *American Journal of Psychiatry*. 2009; 166(9):1048–1054. doi:appi.ajp.2009.08121849[pil]10.1176/appi.ajp.2009.08121849. [PubMed: 19570932]
- Stringaris A, Maughan B, Goodman R. What's in a disruptive disorder? Temperamental antecedents of oppositional defiant disorder: findings from the Avon longitudinal study. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2010; 49(5):474–483. [PubMed: 20431467]
- Swets JA, Dawes RM, Monahan J. Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest*. 2000; 1(1):1–26. DOI: 10.1111/1529-1006.001 [PubMed: 26151979]
- Timbremont B, Braet C, Dreessen L. Assessing Depression in Youth: Relation Between the Children's Depression Inventory and a Structured Interview. *Journal of Clinical Child and Adolescent Psychology*. 2004; 33(1):149–157. DOI: 10.1207/S15374424JCCP3301_14 [PubMed: 15028549]
- van Beek Y, Hessen DJ, Hutteman R, Verhulst EE, van Leuven M. Age and gender differences in depression across adolescence: Real or 'bias'? *Journal of Child Psychology and Psychiatry*. 2012; 53(9):973–985. DOI: 10.1111/j.1469-7610.2012.02553.x [PubMed: 22512614]
- Van Meter AR, You DS, Halverson T, Youngstrom EA, Birmaher B, Fristad MA, et al. Lams Group. Diagnostic efficiency of caregiver report on the SCARED for identifying youth anxiety disorders in outpatient settings. *Journal of Clinical Child and Adolescent Psychology*. 2016; :1–15. DOI: 10.1080/15374416.2016.1188698
- Venkatraman ES. A permutation test to compare receiver operating characteristic curves. *Biometrics*. 2000; 56(4):1134–1138. [PubMed: 11129471]
- Weissman MM, Wickramaratne P, Nomura Y, Warner V, Pilowsky D, Verdelli H. Offspring of depressed parents: 20 years later. *American Journal of Psychiatry*. 2006; 163(6):1001–1008. DOI: 10.1176/ajp.2006.163.6.1001 [PubMed: 16741200]

- Youngstrom EA. A primer on Receiver Operating Characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*. 2014; 39:204–221. DOI: 10.1093/jpepsy/jst062 [PubMed: 23965298]
- Youngstrom EA, Choukas-Bradley S, Calhoun CD, Jensen-Doss A. Clinical guide to the evidence-based assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice*. 2015; 22(1):20–35. DOI: 10.1016/j.cbpra.2013.12.005
- Youngstrom EA, Findling RL, Calabrese JR, Gracious BL, Demeter C, Bedoya DD, Price M. 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 and Adolescent Psychiatry*. 2004; 43(7):847–858. DOI: 10.1097/01.chi.0000125091.35109.1e [PubMed: 15213586]
- Youngstrom EA, Frazier TW, Demeter C, Calabrese JR, Findling RL. Developing a 10-item mania scale from the Parent General Behavior Inventory for children and adolescents. *Journal of Clinical Psychiatry*. 2008; 69(5):831–839. [PubMed: 18452343]
- Youngstrom EA, Genzlinger JE, Egerton GA, Van Meter AR. Multivariate meta-analysis of the discriminative validity of caregiver, youth, and teacher rating scales for pediatric bipolar disorder: Mother knows best about mania. *Archives of Scientific Psychology*. 2015; 3(1):112–137. 1037/arc0000024.supp (Supplemental) 3886/ICPSR36245.v1 (Supplemental). DOI: 10.1037/arc0000024
- Youngstrom EA, Joseph MF, Greene J. Comparing the psychometric properties of multiple teacher report instruments as predictors of bipolar disorder in children and adolescents. *Journal of Clinical Psychology*. 2008; 64(4):382–401. DOI: 10.1002/jclp.20462 [PubMed: 18300293]
- Youngstrom EA, Meyers O, Demeter C, Youngstrom J, Morello L, Piiparinen R, et al. Findling RL. Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders*. 2005; 7(6):507–517. DOI: 10.1111/j.1399-5618.2005.00269.x [PubMed: 16403176]
- Youngstrom EA, Meyers O, Youngstrom JK, Calabrese JR, Findling RL. Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry*. 2006; 60(9):1013–1019. DOI: 10.1016/j.biopsych.2006.06.023 [PubMed: 17056395]

Figure 1.

Receiver operating characteristic analyses comparing baseline CASI-4R depression symptom severity and symptom count caregiver reports, and CASI-4R depression symptom severity and symptom count teacher reports. AUC = Area Under the Curve.

Figure 2.

Nomogram example using the CASI-4R depression subscales.

Note. Jacob is a 10-year old Hispanic boy referred to the clinic due to irritable and sulky mood, disruptive and oppositional behavior, and lack of motivation in school. According to his mother, she previously experienced depression. She completed the CASI-4R depression and dysthymia subscales for Jacob, and her responses yielded a score of 17.

Steps to use the nomogram (adapted from Straus et al., 2011):

1. Select the pretest probability and plot it on the first line on the left side of the nomogram. In this example, we use 17% as the base rate of depressive disorders in an outpatient clinic.
2. Determine and plot the diagnostic likelihood ratio (DLR) associated with the risk factor on the second line. For instance, we found that depression in parents is associated with a three-fold increase in risk for developing major depression (DLR = 3.0; Weissman et al., 2006).
3. Connect the dots and extend the line to the third line to determine that the posttest probability estimate is 38%.
4. To incorporate additional information, repeat the process using the posttest probability as the new pretest probability estimate in the second nomogram.
5. Plot the DLR associated with the parent's CASI depression severity score; in this example, a score of 17 corresponds to a DLR of 3.73 (see Table 3).
6. Connect the dots to determine the new posttest probability estimate of 70%.

Table 1
Descriptive Statistics for Clinical and Demographic Variables, and Bivariate Tests of Association with Narrow Mood Disorder Status

Variable	Any Narrow Mood (<i>n</i> = 63) <i>M</i> (<i>SD</i>) or <i>n</i> (%)	No Narrow Mood (<i>n</i> = 622) <i>M</i> (<i>SD</i>) or <i>n</i> (%)	Effect Size
Age	10.08 *** (1.58)	9.29 (1.92) *	<i>d</i> = 0.41
Female	24 (36%)	202 (32%)	phi = 0.03
Race (White %)	44 (67%)	406 (64%)	phi = 0.02
Axis I Diagnoses at Baseline	3.33 *** (1.39)	2.41 (1.24)	<i>d</i> = 0.72
CASI-4R Symptom Severity - Caregiver	13.67 *** (6.23)	8.85 (5.48)	<i>d</i> = 0.84
CASI-4R Symptom Count - Caregiver	3.78 *** (2.74) *	2.13 (2.25)	<i>d</i> = 0.70
CASI-4R Symptom Severity - Teacher	7.15 * (4.59)	5.44 (4.47)	<i>d</i> = 0.38
CASI-4R Symptom Count - Teacher	1.87 * (1.80)	1.21 (1.78)	<i>d</i> = 0.37

Note: Where data points were missing, effect sizes were calculated out of total number of available cases. Teacher report included 42 cases with any narrow mood, and 418 cases with no narrow mood.

* $p < .05$,

** $p < .005$,

*** $p < .001$, two-tailed, based on *t*-test (means), Levene's test (*SD*/variance), or χ^2 (proportions); effect sizes are Cohen's *d* (means) or phi (proportions).

Table 2
Areas Under the Curve from Receiver Operating Characteristic Analyses for Narrow
Mood Disorder Status

Index Test	Area Under Curve	95% Confidence Interval		Difference between AUCs
		Lower	Upper	
CASI-4R Symptom Severity - Caregiver	.79 ^{***}	.75	.83	$p = .001$
CASI-4R Symptom Count - Caregiver	.78 ^{***}	.74	.83	
CASI-4R Symptom Severity – Teacher	.56	.50	.62	$p < .001$
CASI-4R Symptom Count – Teacher	.60 [*]	.54	.66	

Note:

^{*}
 $p < .05$,

^{**}
 $p < .005$,

^{***}
 $p < .001$, two tailed. Caregiver report was always significantly more discriminating than any teacher report, as shown by the confidence intervals.

Table 3
Different CASI-4R Caregiver Report Severity Scoring Optimal Thresholds and Multi-level Diagnostic Likelihood Ratios (DLRs) for Narrow Mood Diagnosis

Cut Score	Sample Coverage	Diagnostic Likelihood Ratio (DLR)
<i>Multilevel DLRs (based on optimal thresholds)</i>		
0 to 7.49	58%	0.18
7.50 and above	43%	2.31
<i>Multilevel DLRs (based on more informative thresholds)</i>		
0 to 7.99	58%	0.18
8 to 11.99	20%	1.55
12 to 14.99	9%	2.24
15+	13%	3.73

Note: Diagnostic likelihood ratios measure the change in the odds of diagnosis associated with a particular score range. DLRs less than 1 indicate a decrease in the odds, whereas DLRs greater than one indicate an increase in odds of diagnosis by that magnitude. Because we pro-rated scores if someone skipped one or two items, it is possible to have non-integer scores and thresholds.